DOCUMENT RESUME

ED 287 282 FL 016 938

AUTHOR	Bailey, Kathleen M., Ed.; And Others
TITLE	Language Testing Research. Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986).
INSTITUTION	Defense Language Inst., Monterey, Calif.
PUB DATE	87
NOTE	187p.; For individual papers, see FL 016 939-947 and ED 267 645.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Adaptive Testing; Audiolingual Skills; Cheating; Computer Assisted Testing; Discourse Analysis; *Language Proficiency; *Language Tests; Oral Language; Reading Comprehension; *Reading Tests; Research Projects; Scoring; Second Language Instruction; Spanish; Statistical Analysis; Student Placement; *Test Format; *Test Interpretation; *Test Items; Writing Evaluation; Writing Skills
IDENTIFIERS	*ACTFL ETS Language Proficiency Guidelines; Rasch Model

ABSTRACT

This collection of 10 selected conference papers report the results of language testing research. Titles and authors are: "Computerized Adaptive Language Testing: A Spanish Placement Exam" (Jerry W. Larson); "Utilizing Rasch Analysis to Detect Cheating on Language Examinations" (Harold S. Madsen); "Scalar Analysis of Composition Ratings" (Grant Henning, Fred Davidson); "The Graphic Representation of Language Competence: Mapping EFL Proficiency Using a Multidimensional Scaling Technique" (Russanne Hozayin); "An Experiment with a Narrative Discourse Test" (Steven Ross); "Some Problems Concerning the Interpretation of Passage Correction Tests" (Terence Odlin); "Performance Profiles for Academic Writing" (Liz Hamp-Lyons); "Investigations into the Appropriate Use of Various Item Types for Testing Specific Levels of Reading Proficiency" (Sandra McIntyre); "Grading Reading Passages According to the ACTFL/ETS/ILR Reading Proficiency Standard: Can It Be Learned?" (Dale L. Lange, Pardee Lowe, Jr.); and "A Study of the Comparability of Speaking Proficiency Interview Ratings across Three Government Language Training Agencies" (John L. D. Clark). (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ACKNOWLEDGEMENTS

The Editors wish to thank the Defense Language Institute and the Monterey Institute of International Studies for co-hosting the 1986 Language Testing Research Colloquium.

We are especially indebted to Jacqueline Champagne, Office of the Provost, and Pamela Taylor, Office of the Dean for the School of East European Languages, Defense Language Institute, for their generous assistance in the preparation of this manuscript. Ms. Taylor's expertise with wordprocessing equipment, her attention to detail, and her tremendous patience have greatly increased the quality of the manuscript.

TABLE OF CONTENTS

Computerized Adaptive Language Testing: A Spanish Placement Exam 1

Jerry W. Larson

Utilizing Rasch Analysis to Detect Cheating on Language Examinations 11

Harold S. Madsen

Scalar Analysis of Composition Ratings 24

Grant Henning and Fred Davidson

The Graphic Representation of Language Competence: Mapping EFL Proficiency Using a Multidimensional Scaling Technique 39

Russanne Hozayin

An Experiment with a Narrative Discourse Test 60

Steven Ross

Some Problems Concerning the Interpretation of Passage Correction Tests 70

Terence Odlin

Performance Profiles for Academic Writing 78

Liz Hamp-Lyons

Investigations Into the Appropriate Use of Various Item Types for Testing Specific Levels of Reading Proficiency 93

Sandra McIntyre

Grading Reading Passages According to the ACTFL/ETS/ILR Reading Proficiency Standard: Can it be Learned? 111

Dale L. Lange and Pardee Lowe, Jr.

A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Language Training Agencies 132

John L. D. Clark

COMPUTERIZED ADAPTIVE LANGUAGE TESTING:
A SPANISH PLACEMENT EXAM

Jerry W. Larson
Brigham Young University

Language Placement Testing

Placement testing is a concern of nearly all foreign language departments in institutions of higher education. Each semester in-coming students must be counseled and placed into courses that are best suited to their needs and abilities; they are often directed to particular courses based solely on the number of classes they have taken in secondary school, with no formal placement measure being used. This placement procedure does not take into account several factors, such as the effectiveness of previous teachers, the specific information covered, extracurricular or other out-of-class exposure to the language, and, in general, how well the student can actually use the language.

Among the reasons for this method of placement may be a lack of efficient and accurate placement tests, or insufficient time required for testing and evaluating substantial numbers of students at the beginning of the semester. Testing each student before placement is difficult, since adequate placement tests traditionally have required a significant amount of time--about two to two and a half hours for a fairly thorough placement exam--and a lot of coordination effort, i.e., arranging testing times and locations, administering tests, scoring answer sheets, recording and reporting the results, and finally counseling the students.

A search of the literature yields comparatively little discussion about the development and use of adequate foreign language placement measures; nevertheless, conversations with colleagues indicate there is considerable concern about this area of testing and the paucity of research and development being done. Given the diversity of student backgrounds, learning styles, and levels of preparation, placement tests are desperately needed that are able to "predict success in one course and discourage or reject placement in another" (Callas, 1981, p.27).

Adaptive Testing

Adaptive testing is based upon item response theory, a statistical theory which assumes that there is an underlying latent ability distribution for any ability. The ability level of each individual taking a given test lies somewhere along the latent ability continuum. An adaptive test is designed to locate an examinee's ability along that continuum, rather than to determine the degree to which an individual's ability differs from the abilities of others taking the same measure, as is the case with conventional tests (Olsen, 1982, p.1). Traditional tests present a common set of items for all students taking the test, while adaptive tests tailor the test item difficulty to match the ability level of the individual.

The concept of adaptive testing has existed for several years. The early version of the Stanford-Binet intelligence test, for example, was an adaptive test. It was administered by a trained examiner who, during the test, selected subsequent test items according to the individual examinee's responses to previous questions. However, when it became necessary to test large numbers of students, the test became less adaptive in nature.

Testing Rationales

Test design has generally been guided by one of two rationales. One is to devise a "peak test," a measure that contains items fairly homogeneous in difficulty and that is able to discriminate quite effectively over a very narrow range of the variable being tested but that discriminates poorly outside that range. The other rationale is to develop a "uniform test," one comprising items that are heterogeneous in difficulty and thus test a wide range of ability but with a much lower level of precision.

A third rationale, however, is the "adaptive test," which "can approximately achieve the high point precision of a peaked test and can extend that high level of precision over the wide range of a uniform test" (McBride, 1979, p.3). From a psychometric viewpoint, adaptive tests have the additional advantage of a small measurement error, since they concentrate on a single ability level, yet they test a wide range of the trait scale. They are also able to yield precise measurement with fewer test items for the candidate to answer and with increased reliability and external validity (McBride, 1979, pp.1-4).

Adaptive Tests and Computers

Adaptive tests can be administered in three ways: 1) by a skilled examiner, 2) using a specially designed paper-and-pencil test, or 3) at a computer terminal. If only a few students are required to take the test, it may be possible to have skilled testers who can administer the test to each person individually. But, if large numbers of students are to be tested, using trained examiners in this way is impractical. Paper-and-pencil type adaptive tests are possible with either small or large groups, although certain disadvantages must be considered. In order for a paper-and-pencil test to function as an adaptive test, it would require extensive, complex branching instructions for the examinee to follow. This may confuse the dimension to be measured with the examinee's ability to follow instructions (McBride, 1979, p.24). Pencil-and-paper adaptive tests are also subject to the problems of test security, reprinting, scoring and reporting results.

Among the benefits peculiar to computerized adaptive tests are the following:

a. Affective benefits to examinees:

(1) Since this type of test is tailored to the individual, the examinee does not have to experience the boredom or frustration commonly associated with taking lengthy tests, which often range in difficulty from extremely simple to very difficult.

(2) The examinee spends less time taking the test, since fewer items will be necessary for evaluating performance levels.

(3) Examinees receive immediate feedback, without having to wait for days for their results.

(4) Examinees are able to work at their own pace.

(5) Computerized adaptive testing relieves the student of having to comply with complex testing instructions, since the computer automatically selects the appropriate branching routines.

b. Administrative benefits:

(1) Because the computer controls the branching and individualization of test item sequencing, fewer test administrators are needed.

(2) Less time is needed to administer each individualized test.

(3) Test security is improved.

(4) In the event that some test items need changing, a simple deletion and addition can be accomplished without having to reprint new test booklets.

(5) Pretesting of potential test items is easily accomplished without having them recorded as part of the examinee's score.

(6) Computerized adaptive testing eliminates the problem of cheating and test compromise, since each test is virtually unique.

(7) The headaches of test storage, handling test booklets and answer sheets (and their costs) are eliminated.

(8) Also alleviated are the time and expense involved in scoring and reporting test results.

Computer Adaptive Language Testing (CALT)

Although computer adaptive testing has existed in some fields during the past few years, it is a relatively new to foreign language testing. CALT appears to be very well suited to various kinds of foreign language testing purposes, e.g., placement, diagnostic, and proficiency testing, yet it has been slow finding its place in foreign language testing programs. Considering the increased accessability and use of computers in language instruction, it appears that the time is right for serious efforts toward developing computerized adaptive foreign language tests.

Seminal development of CALT placement and proficiency tests is starting to surface at universities, testing services, and other organizations involved with foreign language teaching and testing. Researchers at Brigham Young University have recently developed CALT placement tests for ESL and Spanish; language testing experts at the University of California, Los Angeles, have produced a CALT test for ESL; the College Board has announced a new computerized placement test for Freshman English programs; and ATFL is supporting development of a CALT proficiency test.

Development of a Spanish Computerized Adaptive Placement Exam

The Department of Spanish and Portuguese at Brigham Young University is faced with placing hundreds of in-coming students in lower division courses each year. Because of the enormity of the task, and considering the time and manpower that would be required, it seemed judicious to utilize the computer to the extent possible to assist in this effort.

Research was undertaken to develop a Spanish computerized adaptive placement exam (S-CAPE). Not only was the computer to serve as the "administrator" of the test but also as the manager, or record keeper, of the test results. Additionally, in its role as test administrator, the computer was to adapt the test to the individual examinee, i.e., present items whose difficulty level matched the ability level of the examinee.

Four principal steps were followed in producing this CALT test: 1) development of a test item bank, 2) determination of a suitable psychometric methodology, 3) selection and development of an appropriate S-CAPE delivery system, and 4) validation and implementation of the test.

Development of the Test Item Bank

A critical step in the development of an effective computerized adaptive placement test is having a substantial number of accurately calibrated test items in the item bank. Since the bulk of the students entering Spanish classes must be placed primarily in one of the first four levels, there must be a sufficient number of items at each of those levels to provide accurate indications of student ability. The task of writing test items to cover the four levels of instruction (Spanish 101, 102, 201 and 202) began in October, 1984. Several competent upper-division and graduate students were recommended by the Department of Spanish and Portuguese to work on the project. Each person was interviewed to determine whether his or her background and experience were suited for the task. Then four research assistants, two native Spanish-speakers and two non-natives, were chosen to help with the project.

The project required items to test grammar, reading, and vocabulary for each of the four levels. 1,443 items were written and given to members of the Spanish faculty, who were asked to judge the level and quality of the questions to the best of their knowledge and ability. (Criteria for "quality" included appropriateness of expression, no misleading distractors, use of authentic language, etc.). Eleven hundred of those items were selected and put into five separate test forms (A-E), each having two parts; these tests were administered to 199 students from the four lower division Spanish courses. The tests were submitted to two types of statistical analyses after students had completed them: conventional and Rasch analyses. The conventional analysis consisted of determining the Hoyt and KR-21 reliability coefficients and the point biserial (PBIS) discrimination coefficients. (PBIS is an item level correlation: right versus wrong on an item correlated with the individual's total score.) The results are presented in Table 1.

Table 1

Reliability and Point Biserial Coefficients of Test Forms A - E

Test Form	Part	Hoyt	KR-21	PBIS
A	1	.95	.93	.79
	2	.94	.92	.82
B	1	.96	.95	.74
	2	.95	.94	.82
C	1	.97	.96	.82
	2	.97	.96	.82
D	1	.96	.95	.79
	2	.95	.94	.83
E	1	.93	.91	.77
	2	.87	.83	.85

n = 199

The Rasch analysis was run using Microscale plus TM, a statistical program written specifically for computing Rasch fit statistics.[1] Using this program, item difficulty indices were obtained for each of the items on each of the five test forms. All of the items were "anchored" together and located along a single difficulty/ability continuum. This enabled us to select items of specified difficulty, which is crucial to constructing an adaptive test. The range of difficulty of the items was from −3.6 to +6.0 logits[2], which represents 96 difficulty levels. The 96 levels were ultimately condensed into 51 levels (0-50), covering a difficulty range of −2.0 to +5.8. This range of difficulty easily spanned the ability level of students in the lower division Spanish courses.

After the test items had been calibrated and assigned a difficulty index, they were again given to members of the Spanish Department for careful scrutiny, checking for errors in spelling, inaccuracies in grammar, misleading or tricky distractors, or any other potential problems present in the items. As a result of this review, an additional 51 items were removed from the item bank, leaving about 1,050 items that could be used in the final placement test.

---

[1] For a discussion of the appropriateness of using the Rasch model for calibrating subtests of various language skills, see Henning, Hudson, and Turner in Language Testing, 2(2) (December, 1985): 141-154.

[2] A logit is a unit of measure along the difficulty continuum.

Each of the remaining items were coded, indicating item type (i.e., grammar, reading, or vocabulary), difficulty index, question number, and the letter of the correct response.[3] Nine items at each of the 51 difficulty levels were included in the final version of the placement test. These test items were placed in a random access file, which was created to provide quick and random selection of items within individual difficulty levels.

Testing Methodology

The test was designed so that the computer would present to each individual examinee items that are in the estimated range of his or her ability. Placement into a specified course would be determined by the student's performance on these items. In order to avoid a student's being given the same item twice, a sequential file was created that contained a key for each question in the test item base. This file was read into an array in memory at the beginning of each individual test session. That array was flagged to indicate whether a question had been used. If the question had appeared previously during the test in progress, it would not be selected a second time.

In order for the test to "zero in" on the ability range of an examinee, the following procedure was used: When a student begins the test, the computer selects an item from difficulty level 20, which is toward the lower end of the ability range of most beginning students. If the examinee answers the item correctly, an item at level 26 is randomly selected and displayed on the screen; if the examinee answers the initial item incorrectly, the next item that appears will be from level 15. For the first six questions, the test will increment the difficulty level by 6 for the next question if the response is correct, or it will decrease by 5 if the response is incorrect. This allows the test to "probe," or advance quickly to the approximate ability level of the examinee. After the first six questions have been attempted, the difficulty level adjusts by plus or minus one, depending on the correctness of the previous response, which allows for pinpointing the examinee's actual level of performance. The test is terminated when one of the following conditions is met: 1) the examinee has five correct responses (not necessarily in succession) at any one difficulty level, or 2) the examinee has four incorrect answers on any one level. It was found that condition #2 was the most common cause for the test terminating, since students met condition #1 only if they were answering at the highest level of difficulty (Level 50). At the conclusion of the test, the computer immediately informs the examinees of their ability level. Using this information, the student refers to a Placement Chart, which indicates the ability levels suggested for each course. These parameters, or cutoff points, were determined from the performance of the students from the courses in question, who originally served as the norming group for the test, and of the students in two validation groups (see below). Students who place near the top of a particular range of course scores are encouraged to enroll in the next higher course, challenging themselves a little more than normally would be required.

---

[3] The items used in this placement test are all objectively-scored multiple-choice items, focusing on either grammar analysis, reading comprehension, or vocabulary recognition.

S-CAPE Delivery System

The S-CAPE (Spanish Computer Adaptive Placement Exam) is designed for administration on a standard IBM-PC, or compatible microcomputer, with 64K memory. Either a monochrome or color monitor may be used. The S-CAPE system requires either two floppy disk drives or one hard disk. With the two-drive system, drive A houses the program disk, and drive B contains the examinee's test performance data. After the test master disk is booted, instructions are displayed, informing the examinees how to proceed with the test[4] followed by instructions to enter their name, social security number, and information about their previous exposure to Spanish; this information is recorded onto each examinee's record file. After the student enters this information, the computer automatically displays the first item of the test. When the examinee has answered the appropriate number of items at his or her level of ability, the test will automatically terminate, and the student's ability level is displayed on the screen.

At the conclusion of an individual test or at the end of the day, the person in charge of the testing operation can display on the screen or print on paper all or selected student records. The floppy data disks will accommodate up to 200 different student data files. (Examinees are not made aware of the procedure for printing the test data files.) The printout of each record contains the examinee's name, social security number, amount of previous Spanish instruction or experience, date the test was taken, time the test began, time the test was completed, total number of correct and incorrect responses, and overall performance level. The printout also displays the response pattern of the examinee, indicating which items (by item identification number) were presented to the examinee, the difficulty index of each item, and whether the item was answered correctly or incorrectly. (This information may prove useful for item response analysis or for isolating particular learning deficiencies of students.) An actual sample printout of one of the students--whose name has been changed--is included in Appendix A.

Implementation of the S-CAPE System

Validation studies were conducted before using the S-CAPE on a large-scale basis. The validation sessions were conducted using students from the courses for which the test was designed. Seventy-four students (ranging in ability from only a brief encounter in high school to students completing Spanish

---

[4] The instructions to the student simply explain that they are to respond by typing the letter of the correct answer choice. All keyboard keys other than those corresponding to answer choices are disabled, thus preventing any disconcerting error messages or other distractions from the task at hand.

302) took the test during the first validation session. This was done during the final week of Fall Semester, 1985, when selected teachers sent to take the placement test students whom they judged definitely ready to move on to the next course level. It was believed that the responses and level designations of these students would provide accurate information for determining the course level cutoff points. However, some students who were not ready to advance to the next higher course inadvertently took the test as well; therefore, it was necessary to review the outcome of the testing session on a student-by-student basis before being able to determine the minimum ability level required for each course.

A second validation session was conducted during Winter Semester, 1986, using 129 students.

Cutoff levels have now been determined specifically for the Department of Spanish and Portuguese at Brigham Young University, which is using the S-CAPE to screen and place students into its courses. Using fifteen microcomputers in the Humanities Learning Resource Center at Brigham Young University, it is possible to test 300-350 students in an eight-hour day, since the tests require on the average approximately 18-20 minutes to complete.[5]

Conclusions and Recommendations

While there is no such thing as a perfect placement test, initial observations and evaluations of the S-CAPE are extremely positive and promising. Further research is recommended, however, in order to produce a test in which one would have near total confidence. For example, additional research might be done to perfect the procedure for determining test termination criteria.

Further research into and application of the computerized adaptive testing format should occur in the areas of proficiency, achievement, and diagnostic testing.

Given the technology of the day, it is possible to expand the scope of computerized adaptive language testing by adding a listening component, using a random-access audio disk player to select listening items that have been properly calibrated. Audio disk technology allows for almost instantaneous retrieval of any prerecorded segments on the disk. The computer could be programmed to seek and retrieve audio segments based upon examinee responses.

The results of the efforts to develop a Spanish computerized adaptive placement test are gratifying. This project, as well as others mentioned above, have shown that it is possible to produce tests that can substantially reduce the burden of language placement testing for language departments with large enrollments.

---

[5] The minimum amount of time spent by an examinee taking the test is six minutes, while the maximum has been 59 minutes.

References

Callas, Dennis. 1981. Assessment data, the academic system and evaluation of the placement process. Community College Review, 9(2), 27-33.

Henning, Grant, Hudson, Thom, and Turner, Jean. 1985. Item response theory and the assumption of unidimensionality for language test. Language Testing, 2(2), 141-154.

McBride, James R. 1979. Adaptive mental testing: the state of the art. Technical Report 423. Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences.

Olsen, James B. 1982. Psychometric recommendations for computerized adaptive testing. Monograph. Orem, Utah: WICAT Systems.

Appendix A

Spanish Computer Adaptive Placement Exam Report

Jane Jones 576-56-0823 Date: 04-02-1986 Time started: 11:56:06

Spanish Background:
One semester in college

Difficulty Level	Test Item	Student Answer	Right/ Wrong
20	174	A	w
15	123	D	r
21	182	C	w
16	134	B	r
22	191	B	r
28	244	A	w
23	202	C	w
22	188	B	r
23	197	A	r
24	212	D	r
25	220	D	w
24	206	B	w
23	198	B	r
24	205	C	r
25	218	A	w
24	210	A	w
23	204	D	r
24	213	B	w
23	196	C	r
24	211	A	w

There were four incorrect answers at level 24
LEVEL: 24 Number right: 23

Test completed at 12:07:52

UTILIZING RASCH ANALYSIS TO DETECT CHEATING ON LANGUAGE EXAMINATIONS

Harold S. Madsen
Brigham Young University

The purpose of this investigation is to assess the effectiveness of the Rasch procedure in conducting appropriateness measurement--notably the detection of cheating on multiple-choice language tests. Of the several statistical models that have been developed to further latent trait analysis related to objective tests and survey-instruments, the Rasch procedure seems to be the most appropriate option for examining the small data samples (say 50 to 100 subjects) often encountered in ESL and FL settings (Lord 1983, Henning 1984).

Unlike multiple-parameter statistical models which incorporate factors such as discrimination and guessing into their formulas, the Rasch one-parameter model is based on the single factor of response probability--the likelihood of each individual's answering a given item correctly and the corollary probability of items being responded to correctly. With the logarithmic calculations built into the Rasch model, it is possible to generate 1) remarkably precise calibrations of item difficulty and examinee performance, 2) item calibrations essentially independent of the sample populations being measured, and 3) measures of aberrant or unexpected behavior, such as the tendency for a difficult item to be answered correctly by some low-proficiency examinees.

Rasch analysis has been applied to a wide variety of testing needs, including item banking (Woods and Baker 1985, Henning 1986), sample-free item analysis (Wright and Masters 1982), and identification of item bias (Hulin, Drasgow and Parsons 1983, 167-170; Madsen and Larson 1986). It has even been used to calibrate speaking test interview items allowing for partially-correct responses (Griffin 1985) and to analyze a scalar self-rating language ability instrument and a writing scale (Davidson and Henning 1985, Henning and Davidson 1986).

Appropriateness Measurement

Another potential Rasch application of considerable interest is that of appropriateness measurement, a broad area of investigation only part of which is concerned with the problem of cheating. Two important recent texts each devote a major chapter to appropriateness measurement--Item Response Theory: Application to Psychological Measurement (Hulin and others 1983, 110-151) and New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing (Weiss 1983, 109-131). The former authors define such research as "any method of detecting individuals for whom test scores fail" as measures of the latent trait, ability, or interest (1983, 111). Investigations in this area examine all sorts of test taking anomalies. These aberrations can be roughly classified into two categories: unintentional behaviors and deliberate tampering with test responses.

Relative to the first category, Wright and Stone (1979, 165-180), Levine and Drasgow (1982), Hulin and others (1983, 112-114) mention several examples of unintentional behaviors that can result in misleading test scores. These include a student's guessing, being inattentive or bored, "fumbling" because of confusion over the intent of test questions or exam directions, stumbling on items because of their cultural bias, not fully completing an exam, mismarking a machine-scorable answer sheet, or simply looking at items (as on an IQ instrument) from a novel or creative point of view. Since he is concerned with the problem of inappropriate score results, Wainer challenges the usual psychometric procedure of point reduction for coping with guessing. He posits that behavior labeled as guessing or "sleeping" (carelessness due to inattentiveness or boredom) could in fact be a result of "unusual patterns of knowledge" or other factors (1983, 64-65). In their pioneering article on appropriateness measurement, Levine and Rubin refer to those with deficient test taking strategies: i.e., individuals inclined toward "excessive guessing, reluctance to answer questions unless absolutely sure, [and] ... behavior suggesting lack of experience with test format" (1979, 270).

Unintentional aberrant responses can also stem from special handicaps: for example, a person with limited language proficiency taking a test designed for native speakers (Oller and Perkins 1978, Levine and Rubin 1979), or a student with hearing deficiency taking an aural comprehension test. In addition, a person with an atypical educational background may be handicapped on a test designed for mainstream students (Drasgow 1982).

Turning to the second category of appropriateness measurement, we note that deliberate altering of results also takes many forms. Students recently have tampered with computer files of tests: a graduate student programmed a computer to give himself right answers and other students incorrect answers; in Queens, New York, students used computers to raise 10,000 grades (Moody 1980), and teachers have even tampered with student answer sheets, filling in correct responses for deleted items (Jensen 1980). There are numerous occasions on which answer sheets have been obtained and sold to some candidates in advance of the test. Examinees have also been known to suppress scores on a minimal proficiency examination such as those administered by the government for military draftees (Olsen 1986). More common is an examinee's copying of a number of items from another student during a test. Responses on attitude measures are also at times modified by the respondent in order to match perceived expectations on the part of those using the instrument.

Recent Research on Cheating

A variety of recent studies have disclosed that cheating is a serious problem in today's schools (Roizen 1982, Singhal and Johnson 1983, Nuss 1984). Others have sought to determine the causes of dishonest behavior in the classroom (Shab 1980, Chapman 1980), and even the personality traits of cheaters (Leming 1980, Antion and Michael 1983, Forsyth, Pope and McMillan 1985). In addition, possible solutions to the cheating problem have received considerable attention (Houston 1977, Urbain 1980, Guttman 1984). A related research concern is that of cheating detection. The following studies focus on this effort.

Utilizing IRT Procedures to Detect Cheating

A variety of conventional procedures for identifying cheating have been employed over the years. One statistically-based procedure utilized to verify suspected cheating focuses on the match of wrong answers between the suspected cheater and his or her "source" plus the wrong-answer patterns of the remaining students (Cody 1985, 136-137). A variation on the procedure utilizes students in adjacent seats as sources and two random papers three or more seats away for the comparison (Houston 1983).

Hulin and others after surveying a variety of traditional or heuristic approaches designed to identify inappropriate test performance (Ghiselli's prediction of predictability, Jacob's weighted average, as well as Donlon and Fischer's personal biserial) find them imprecise and largely ineffective in identifying normal and aberrant behavior patterns (1983, 114-121).

On the other hand, recent IRT research in the area of appropriateness measurement has yielded very impressive results in identifying both normal response patterns and aberrant patterns. Forerunners in this movement, such as Levine and Rubin, repeatedly assure readers that they are not committed to the indices they have developed (1979, 286). But there tends to be some reservation among these scholars about the effectiveness of the one-parameter model: For example, in his dissertation research, Drasgow found the three-parameter model superior to the one-parameter model in distinguishing between normal and aberrant response patterns on GRE multiple-choice Verbal Aptitude items (1978).

But Drasgow does stop short of advocating the exclusive use of a three-parameter model, particularly if cost is a factor (1982, 305-307). Moreover, Hulin and others indicate that differences between the three- and one-parameter models are not sufficient to warrant the use of the former model "under all circumstances" (1983, 141). Given the suitability of the one-parameter model for small-sample test studies, it was decided to undertake our appropriateness research by utilizing the Rasch model.

Studies on Cheating

Studies Involving TOEFL Cheaters

Since cheating on exams is such a widespread problem, it was enticing to determine whether principles of appropriateness measurement could be utilized as a kind of "radar" to detect such aberrant performance. At the outset, however, it was obvious that our "radar" could not be expected to detect all dishonest performance on a language exam: While it should be possible to detect a weak student's copying a page or two from a much superior student, it wouldn't be possible to detect an advanced language learner copying from another advanced student (Levine and Rubin 1979, 270), because their pattern of expected responses would be very similar. Moreover, if Student A took an exam on behalf of Student B, it would be possible to detect this only if Student A performed abnormally, for example by arbitrarily suppressing his score.

Being aware that Educational Testing Service had recently apprehended some cheaters on the TOEFL--individuals who had received payment to take the exam for foreign candidates seeking admission to American universities--the author arranged to attempt their identification by means of Rasch analysis. It was hypothesized that to escape detection these professional cheaters would suppress their scores. Arbitrarily missed items would then flag the cheater as an outlier on the computer printout--an able student missing a number of relatively easy items. While the initial investigations failed to identify cheaters, they did set the stage for follow-up work which was successful in spotting the aberrant behavior of persons tampering with test items.

TOEFL Study One. The first study involved two known cheaters on the TOEFL (Test of English as a Foreign Language). Two sets of answer sheets, with identification removed, were obtained from Educational Testing Service, the agency responsible for preparing and administering TOEFL. Selected from exam papers at five centers, there were 29 test papers from TOEFL Form 7; and there were 30 from TOEFL Form 15. Each of the two sets of answer sheets included the paper of one unidentified cheater.

The two groups were Rasch analyzed separately utilizing the Microscale Plus routine developed by Wright and Linacre (1985) for the IBM personal computer. The outlier-sensitive fit statistic (or "outfit") was obtained, since this identifies unexpected events, such as examinees performing in an aberrant fashion. But contrary to expectations, no clearly identifiable outliers were identified (Figure 1). In short, the cheaters had not been detected. The reason for this became apparent, however, after their ID numbers were provided by ETS: They had performed so well (with scores of 610 and 623) that it was apparent they could not have suppressed their scores.

TOEFL Study Two. ETS next agreed to send two additional groups of papers (identification removed as before), with cheaters whose scores were well under 600: one with a score of 557 and the other with a score of 563. The paper of one cheater was included in a group of 46 answer sheets from TOEFL Form 8; the paper of the second cheater was included in a batch of 44 answer sheets from TOEFL Form 9. As in the first TOEFL study, data from five different testing centers were utilized.

Again the outfit plots (Figure 2) generated by Rasch analysis of these second two groups failed to identify the two known cheaters. Interestingly, however, the analysis did identify an outlier with a dramatically high outfit score. For such a surprising outfit measure, this weak student (number 13) would need to exceed the chance probability of answering correctly a number of items well beyond his level of ability. The most plausible explanation was that he had cheated. But his low score informs us that this would not have been the result of score suppression (the object of this investigation). This probable cheater was reported to TOEFL officials.

Our initial investigation, then, into the response patterns of known cheaters reveals that either they had not attempted to suppress their scores, or that such an effort had been so minimal that the analysis was unable to detect it.

Cheating Studies Utilizing Artificial Data

Contrived Cheater Study One. This investigation attempted to determine how effective Rasch analysis would be in identifying a small percentage of manipulated scores. Drawing upon data from 183 ESL students, the author selected two papers at the low end of the scale and rescored them so that correct answers were given to five items in the top quartile of difficulty (see Madsen and Larson 1986, 59). When the data were Rasch analyzed again, the outfit scores for these two papers had jumped dramatically: from approximately 1 to well over 4 -- the highest outliers of the entire group of 183 students (Figure 3).

Contrived Cheater Study Two. In contrast with this spuriously high score situation, it was next decided to devise a more realistic setting, one parallel to that hypothesized for professional cheaters on the TOEFL. The cheater's paper from the TOEFL Form 7 group was selected, and a "miss" of one in every ten responses was manipulated. (These incorrect responses plus those missed as a matter of course by this non-native cheater resulted in a TOEFL score of 577). The data were again Rasch analyzed. Now with a suppressed score, the paper had the highest outfit of any in the group (Figure 4).

But while such an outfit score could help corroborate allegations of cheating, it might not always flag decidedly aberrant test behavior. Like Levine and Rubin (1979), we needed to evaluate a situation well beyond the 10 percent level of manipulation. After examining 4, 10, 20 and 40 percent score modifications of artificial data, they had found that only when tampering reached the 20 percent level or higher did they achieve good detection rates.

Contrived Cheater Study Three. Now a TOEFL scenario was devised, involving a supposed native speaker cheater with a 20 percent score suppression. This contrived test paper was added as examinee number 30 in the TOEFL Form 7 group. The data were again Rasch analyzed, this time with impressive results. The native-language "cheater" had an outfit score approximately three times as great as the highest regular student (Figure 5).

Discussion

The appropriateness-measurement investigations reported here demonstrate that the Rasch one-parameter model is able to detect some aberrations in test responses due to cheating. The TOEFL studies remind us, however, of the limitations of IRT "radar": Unless spuriously high or spuriously low scores are generated, the cheater will go undetected. The studies involving score manipulation point out a second limitation: Unless the score suppression or inflated scores due to copying or memorizing a block of answers are substantially above the 10 percent level, there is a danger of "false arrest." But as Levine and Rubin found with another model (1979), when the alterations approach the 20 percent level or beyond, Rasch analysis can easily detect aberrant performance. Caution is needed, of course, in attributing all extreme outlier or aberrant performance to cheating. As we have seen, inappropriate performance on a test could result from a variety of causes.

The door is now open for testing organizations, educational institutions, and even a great many individual teachers to deal more thoroughly and systematically than ever before with faulty measurement on language exams. Not only are we more able to apprehend cheaters but through IRT procedures such as Rasch analysis we can help rectify inequities in the evaluation process. As valuable as the new technology is, however, it will need to be accompanied by human investigation into the causes of the aberrant performance.

References

Antion, David L. and William B. Michael; 1983. Short-term predictive validity of demographic, affective, personal, and cognitive variables in relation to two criterion measures of cheating behaviors. Educational and Psychological Measurement 43(2): 467-482.

Chapman, William; 1980. A cheating problem in Japan. Today's Education 69(4): 42.

Cody, Ronald P; 1985. Statistical analysis of examinations to detect cheating. Journal of Medical Education 60(2): 136-137.

Davidson, Fred and Grant Henning; 1985. A self-rating scale of English difficulty: Rasch Scalar Analysis. Language Testing 2:2 (December), 164-179.

Drasgow, Fritz; 1978. Statistical Indices of the appropriateness of aptitude test scores. Unpublished Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Drasgow, Fritz; 1982. Choice of test model for appropriateness measurement. Applied Psychological Measurement 6:3 (Summer), 297-308.

Forsyth, Donelson R., William R. Pope, and James H. McMillan; 1985. Students' reactions after cheating: an attributional analysis. Contemporary Educational Psychology (1): 72-82.

Griffin, Patrick E.; 1985. The use of Latent Trait Models in the calibration of tests of spoken language in large-scale selection-placement programs, 149-161. New Directions in Language Testing. Y. P. Lee, Angela C.Y.Y. Fok, Robert Lord and Graham Low (Eds.) Oxford: Pergamon Press.

Guttman, Joseph; 1984. Cognitive morality and cheating behavior in religious and secular school children. The Journal of Educational Research 77(4): 249-254.

Henning, Grant; 1984. Advantages of Latent Trait Measurement in language testing. Language Testing 1(2): 123-133.

Henning, Grant; 1986. Item banking via DBASE II: The UCLA ESL Proficiency Examination experience. Technology and Language Testing. Charles W. Stansfield (Ed.) Washington, D.C.: TESOL.

Henning, Grant and Fred Davidson; 1986. Scalar analysis of composition ratings. Paper presented at the Eighth Language Testing Research Colloquium, Monterey, Ca.; February 28 (This volume).

Houston, John P; 1977. Cheating behavior, anticipated success-failure, confidence, and test importance. Journal of Educational Psychology 69(1): 55-60.

Houston, John P.; 1983. Alternate test forms as a means of reducing multiple-choice answer copying in the classroom. Journal of Educational Psychology 75(4): 572-575.

Hulin, Charles L., Fritz Drasgow, Charles K. Parsons; 1983. Item Response Theory: Applications to Psychological Measurement. The Dorsey Professional Series. Homewood, Illinois: Dow Jones-Irwin.

Jensen, A.R.; 1980. Bias in Mental Testing. New York: Free Press. Reported in Hulin, Charles L., Fritz Drasgow, and Charles K. Parsons; 1983. Item Response theory: Application to Psychological Measurement. Homewood, Illinois: Dow Jones-Irwin.

Leming, James S.; 1980. Cheating behavior, subject variables, and components of the internal-external scale under high and low risk conditions. Journal of Educational Research 74(2): 83-87.

Levine, Michael V. and Fritz Drasgow; 1982. Appropriateness measurement: review, critique and validating studies. British Journal of Mathematical and Statistical Psychology 35:42-56.

Levine, Michael V. and D.F. Rubin; 1979. Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics. 4(4): 269-290.

Lord, Frederic M.; 1983. Small N justifies Rasch Model, 51-61. New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing, David J. Weiss (Ed.) New York: Academic Press.

Madsen, Harold S. and Jerry W. Larson; 1986. Computerized Rasch Analysis of item bias in ESL tests, 47-67. Technology and Language Testing, Charles W. Stansfield (Ed.) Washington, D.C.: TESOL

Moody, Sid; 1980. Cheating: a national pastime. Today's Education 69(4): 32.

Nuss, Elizabeth M.; 1984. Academic integrity: comparing faculty and student attitudes. Improving College and University Teaching 32(3): 140-144.

Oller, John W., Jr. and Kyle Perkins; 1978. Language in Education: Testing the Tests. Rowley, Massachusetts: Newbury House Publishers, Inc.

Olsen, James. Personal communication, 1986.

Roizen, M.A.; 1982. The population explosion: test-taking strategies project. School of Education, Hebrew University of Jerusalem, course paper. Reported in Cohen, Andrew D. On Taking Language Tests: What the Students Report. Language Testing 1(1): 70-81.

Schab, Fred. 1980. Cheating in high school: differences between the sexes (revisited). Adolescence 15(Winter): 959-965.

Singhal, Avinash C. and Patti Johnson; 1983. How to halt student dishonesty. College Student Journal 17(1): 13-19.

Urbain, Kermit; 1980. Trust them not to cheat. Today's Education 69 (4):35

Wainer, Howard; 1983. Are we correcting for guessing in the wrong direction? 63-80. New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing, David J. Weiss (Ed.) New York: Academic Press.

Weiss, David J. (Ed.); 1983. New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. New York: Academic Press.

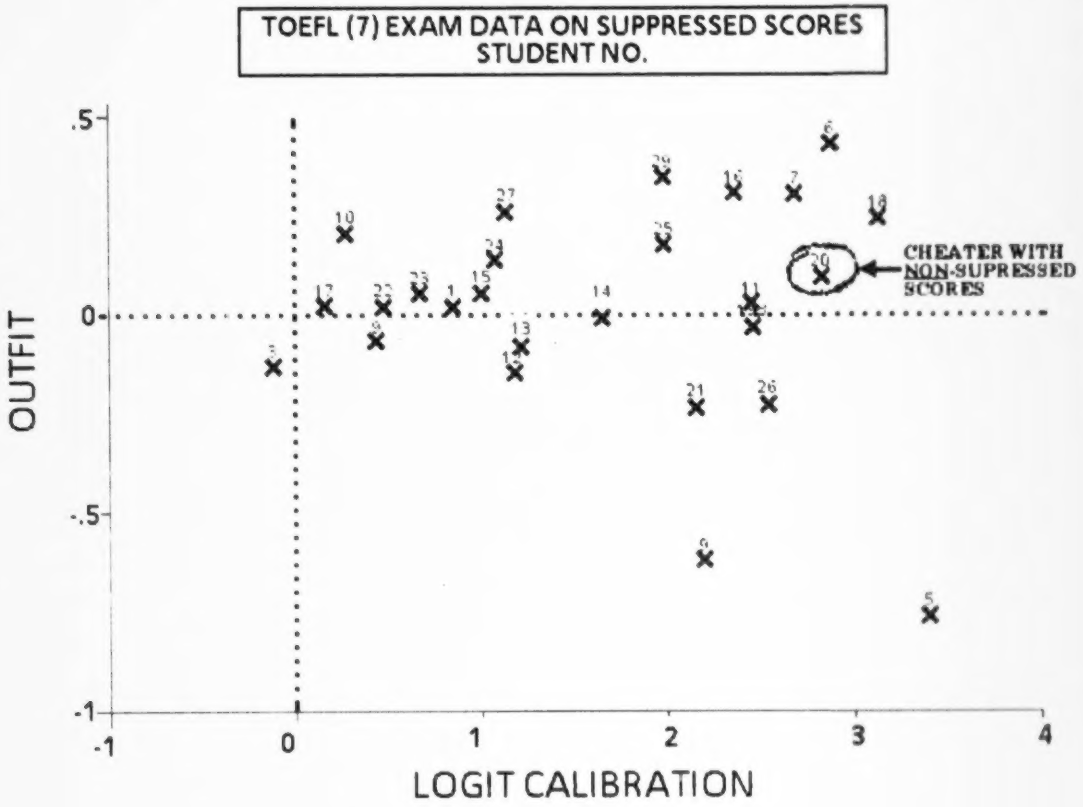
Woods, Anthony and Rosemary Baker; 1985. Item Response Theory. Language Testing 2:2 (December), 119-140.

Wright, Benjamin D. and Michael Linacre; 1985. Microscale Manual. Westport, Ct.: Mediax Interactive Technologies.

Wright, Benjamin D. and Geofferey N. Masters; 1982. Rating Scale Analysis: Rasch Measurement. Chicago: Mesa Press.

Wright, Benjamin D. and Mark H. Stone; 1979. Best Test Design. Chicago: Mesa Press.

Figure 1


TOEFL (7) EXAM DATA ON SUPPRESSED SCORES
STUDENT NO.

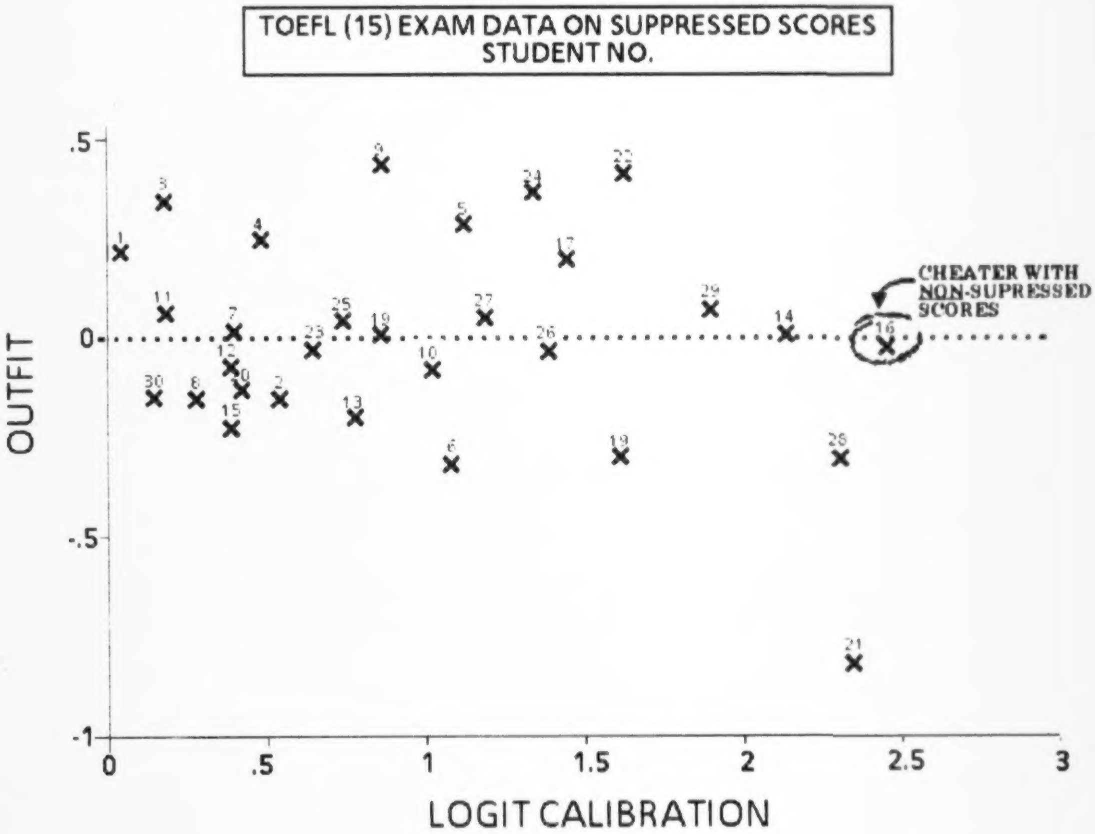
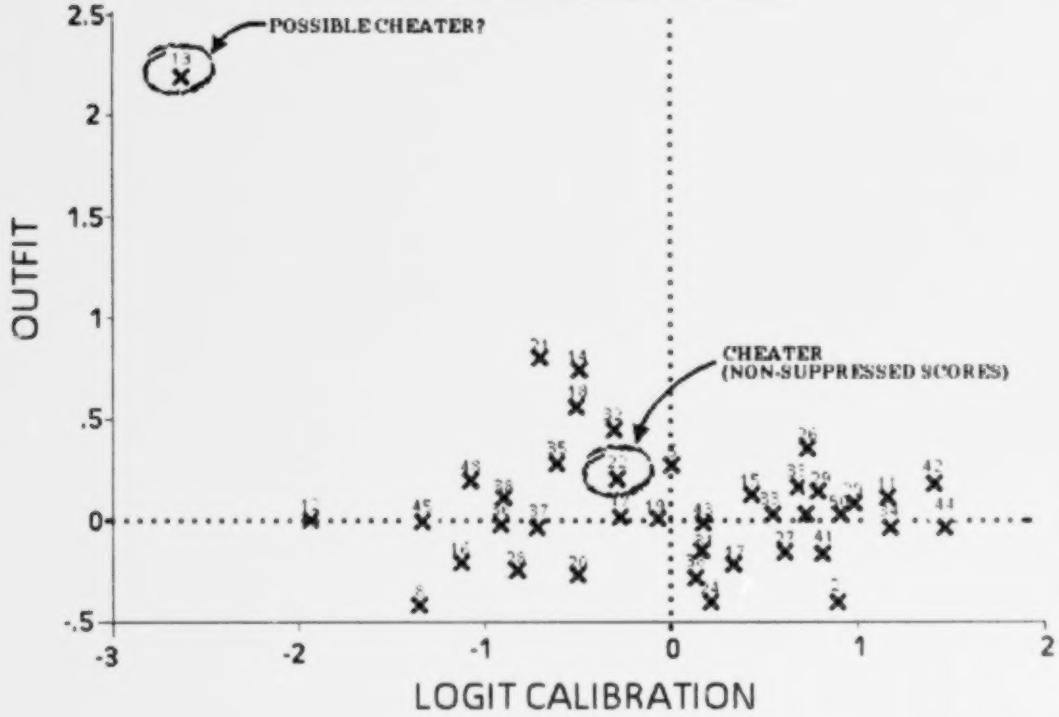
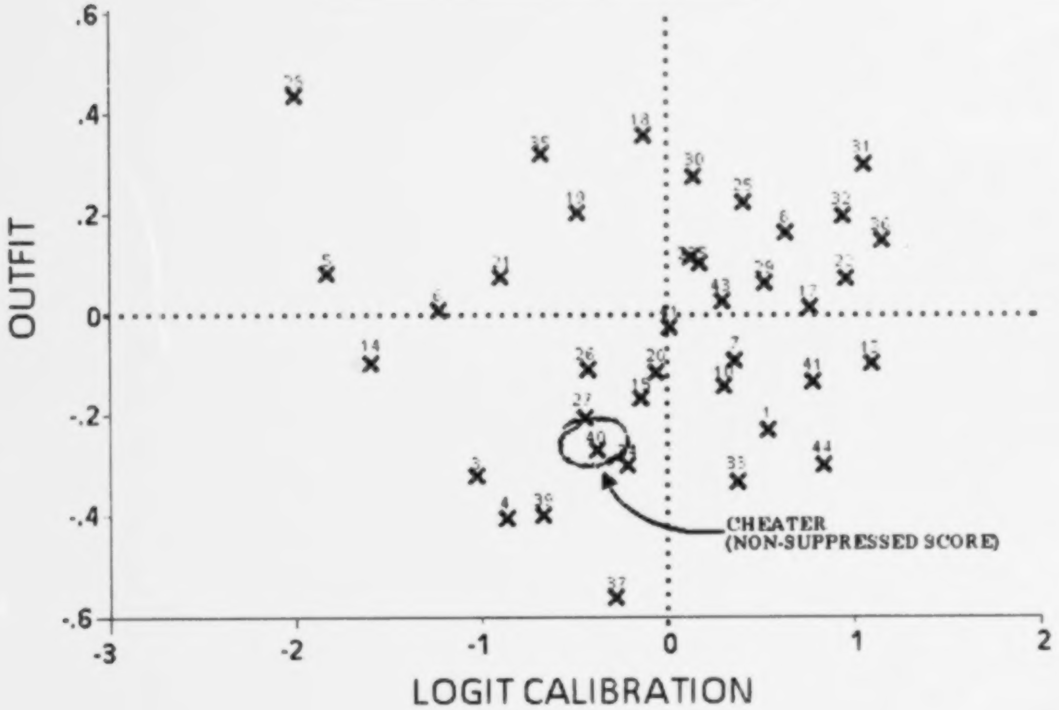
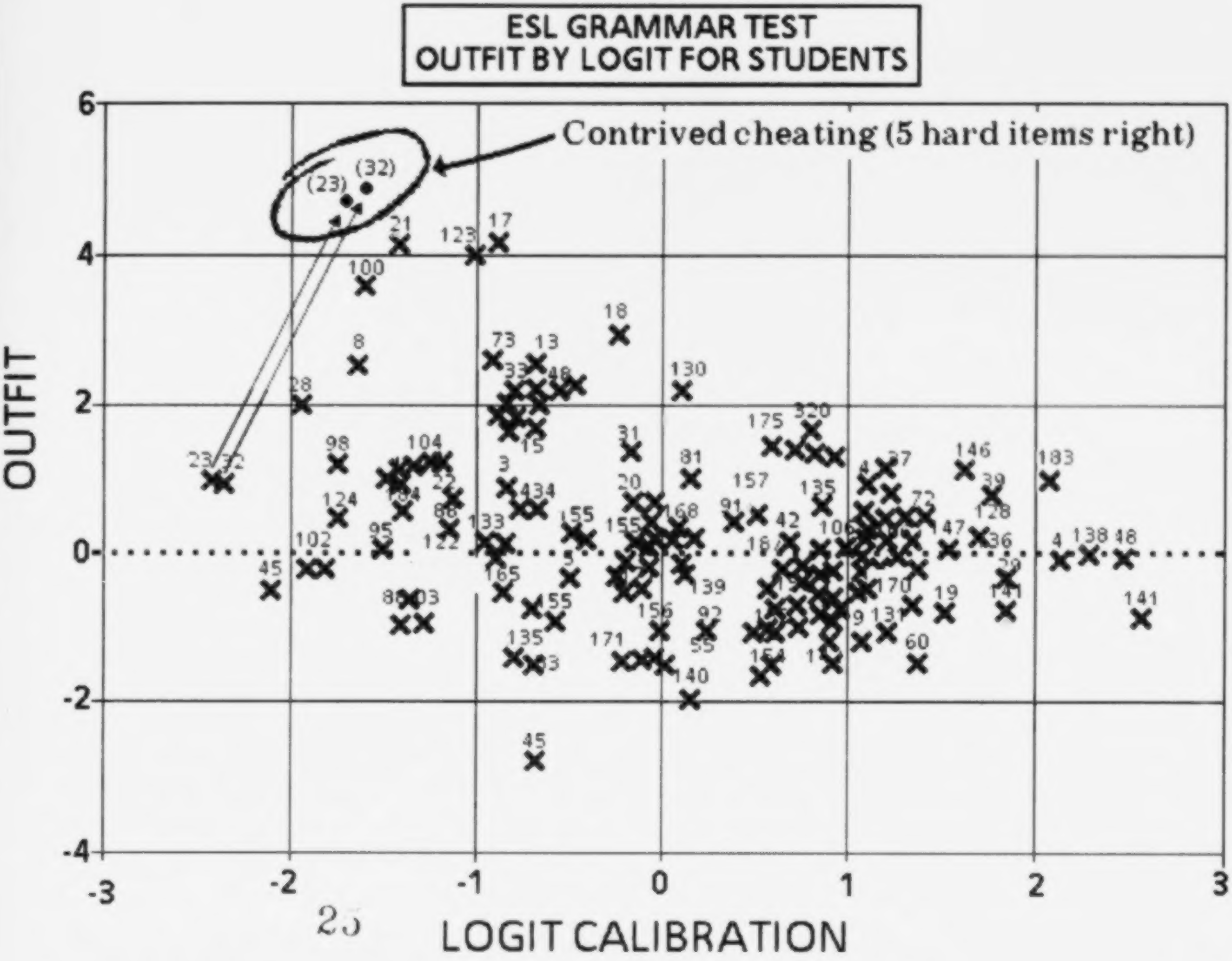
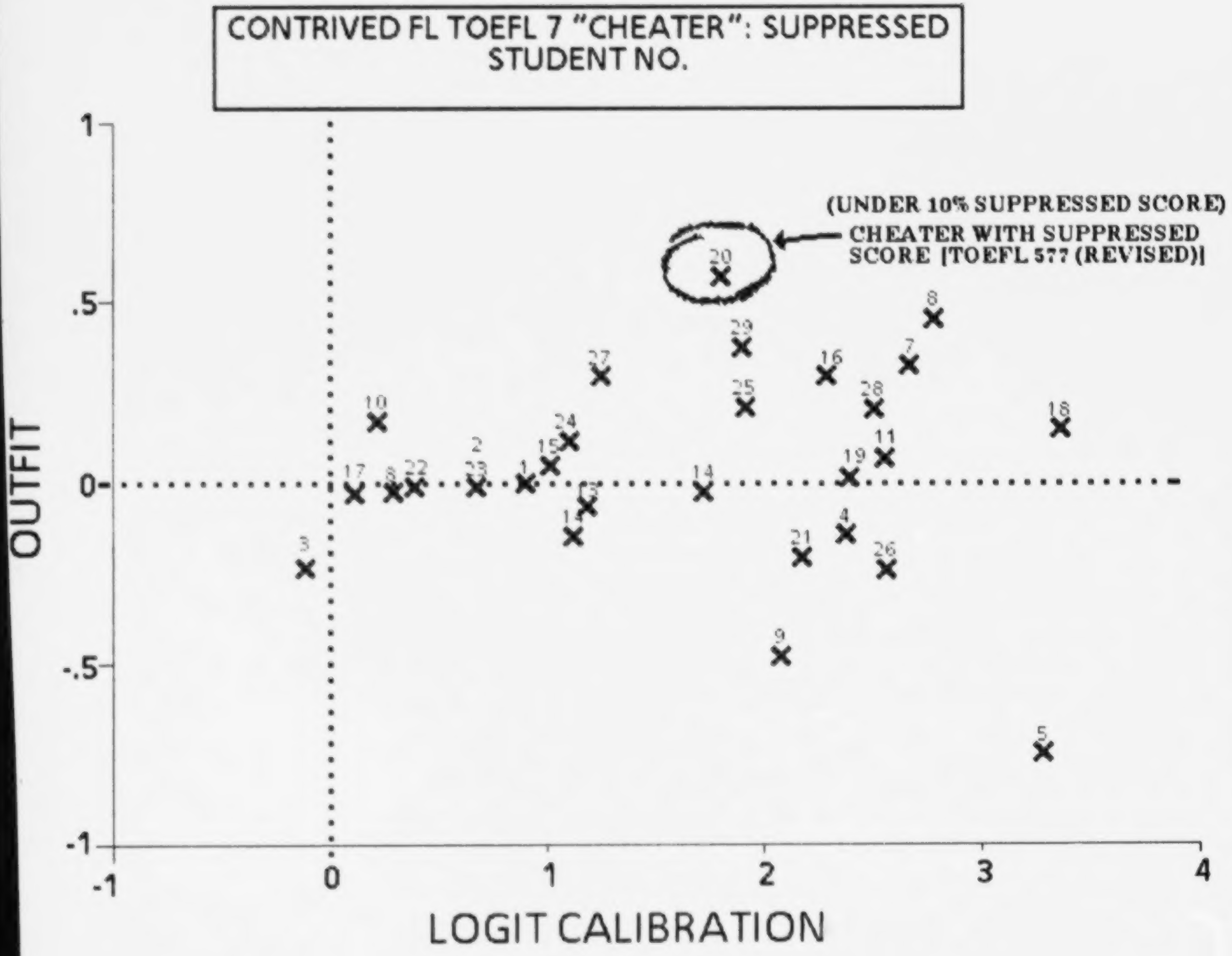

TOEFL (15) EXAM DATA ON SUPPRESSED SCORES
STUDENT NO.

Figure 2

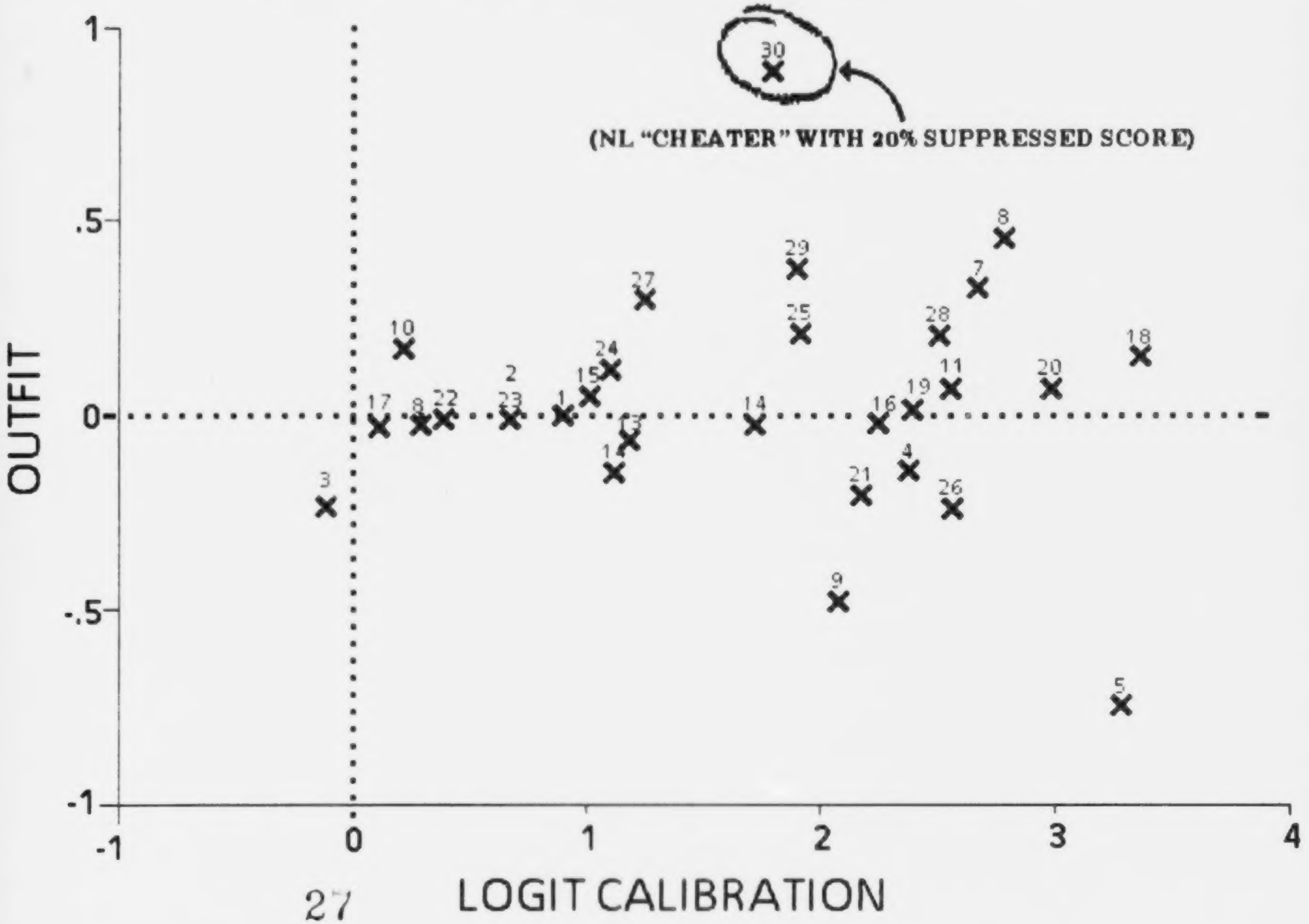

FL "CHEATER(S)" ON TOEFL 8
STUDENT NO.


FL "CHEATER(S)" ON TOEFL 9
STUDENT NO.


ESL GRAMMAR TEST
OUTFIT BY LOGIT FOR STUDENTS
Contrived cheating (5 hard items right)
(23)
(32)
OUTFIT
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
LOGIT CALIBRATION
23
32
21
123
17
100
18
8
73
13
130
28
175
320
98
104
31
81
146
183
3
157
135
37
39
124
20
168
91
72
128
155
95
102
133
42
147
36
4
138
48
45
122
165
5
139
170
19
141
156
92
131
135
171
55
60
140
45


CONTRIVED FL TOEFL 7 "CHEATER": SUPPRESSED STUDENT NO.
(UNDER 10% SUPPRESSED SCORE)
CHEATER WITH SUPPRESSED SCORE [TOEFL 577 (REVISED)]
OUTFIT
LOGIT CALIBRATION
1
.5
0
-.5
-1
-1
0
1
2
3
4
20
29
8
7
16
27
25
28
10
18
24
2
11
15
19
17
6
22
23
1
14
13
12
4
21
26
3
9
5

CONTRIVED NL TOEFL 7 "CHEATER": SUPPRESSED STUDENT NO.


(NL "CHEATER" WITH 20% SUPPRESSED SCORE)

SCALAR ANALYSIS OF COMPOSITION RATINGS

Grant Henning
Fred Davidson
University of California, Los Angeles.

Introduction

Inasmuch as there is reliance on various kinds of rating scales in evaluating language learner performance in writing and speaking, it seems appropriate to investigate efficacy of such scales for use with particular examinees for particular decision-making purposes. Classical analyses of language performance scales have usually been limited to considerations of interrater reliability and of concurrent, predictive, or construct validity (Henning, 1983, 1984).

In general, additional information is needed regarding the accuracy and incremental nature of language performance scales at all points along the scoring continua which they define. Further information is required about the interrelatedness and appropriate combinatory weightings of oral performance categories such as fluency and pronunciation accuracy, and of written performance categories such as content and mechanics. Measures of response validity are needed to enable judgments regarding the particular persons, performance categories, and performance levels for which fair measurement decisions can be made. Performance scales should also be evaluated in terms of the consistent match of the difficulty of performance they are purported to gauge to the ability range of the examinees with whom the scales are used. Reliability of scales should be estimated--not merely as a global interrater correlation coefficient--but also as an index of the sensitivity of the scale to differentiate strength of person performance at all levels of the performance continuum. These needs in the evaluation of performance scales have suggested the areas of investigation for the present study of a composition rating scale currently in use to evaluate writing at all levels of an English-as-second-language (ESL) instructional program at the University of California,Los Angeles (see Appendix A). Specifically, the following questions are being investigated with regard to the performance scale considered:

1. What is the comparative difficulty of the subscale categories (Content, Organization, Expression, Structure, and Mechanics)?

2. How are subscale categories interrelated and, thus, what combinatory weightings can be suggested to produce a total score that is most predictive of performance variance?

3. How are subscale categories and performance levels incrementally arranged on a probabilistic writing performance continuum (with regard to equality or non-equality of intervals)?

4. What estimates of scalar reliability can be derived, both as a global measure of person separability, and with regard to the error associated with each estimate of difficulty and ability?

5. How do persons, categories, and performance levels fit the probabilistic predictions of the Rasch Model, thereby exhibiting a kind of response validity?

6. What are the characteristics of the writing of misfitting persons, i.e., persons for whom ratings should not be interpreted with the same confidence in decision making?

These questions along with several related questions are addressed in the present study of a given composition grading scale as it was employed in the rating of compositions produced by students in the advanced levels of the ESL Service Courses at UCLA.

Method

Subjects

143 Subjects applied to enter the UCLA ESL writing courses at the beginning of Fall Quarter, 1985. Any student seeking entry was required to take a composition placement test. The results of the test (in addition to other enrollment criteria) determined eligibility for one of three levels of ESL writing courses. Generally, these 143 students were at a high level of ESL proficiency (above 550 on the TOEFL). The three levels of writing courses span the non-native speaker (NNS)/native speaker (NS) transition at the freshman level of college writing. That is, the lowest course, English 35, was intended for NNSs who do not quite exhibit the command of written English necessary for coping with a first-year writing load. The next course, English 36, is equivalent to first year college writing for native speakers. The final course, English106J, is intended for students who still require refinement of written skills beyond English 36 or who for various reasons wish to study more advanced ESL composition. Of the 143 students who took the test, and on whom this paper reports, 10 were placed in English 35, 87 in English 36 and 17 in English 106J. The remainder were placed on a waiting list for subsequent terms.

Instrumentation

All subjects wrote for 75 minutes on the first day of class. Administration, including instructions, completion of an information sheet, a brief error-detection task, and the 75-minute composition test itself were all designed to require exactly two hours--the normal writing class period. The students were advised to check with the ESL office prior to the second class meeting so that they could determine which course level they were assigned. The composition topic was selected and approved by instructors of the ESL writing courses prior to the exam.

Rating Procedure

Following the administration on the first class day, the ESL writing instructors met with several other experienced ESL writing teachers to score the compositions. In all, nine raters participated. Raters were paid for the scoring task, over and above their normal instructor salaries. Their first task was to participate in a norming session to ensure uniform interpretation and application of the composition rating scale. The actual scale employed along with a set of related performance descriptors is provided in the appendices. (See Appendix A.)

Five papers were chosen at random, xeroxed, and scored by all raters. Then the raters met, shared their grades, and agreed on uniform standards and interpretation of the scoring terminology. It was decided that for placement purposes the five norming papers would be assigned the average score of all the ratings derived during the norming session, but that they would not be included in the study because they represented "pre-norming" ratings.

Following the norming, the 138 remaining papers were scored using the composition rating scale. The rating of these compositions employed the following established procedure. First, each rater took approximately 10-15 papers and read them independently. On finishing that set, the rater returned the set to a "second-rater-needed" box. Each rater then read another set from the unrated pile or from the second rater box. Each rater marked each completed composition with a code number to guarantee that no rater read the same paper twice. Furthermore, the score sheets were kept folded to ensure that no rater saw the score assigned by any previous rater.

If the first and second raters were four or more total points apart on a given composition (on the combined scale of 25), then a supervisor placed that essay into a "third-rater-needed" box. Once that box contained compositions, any returning rater had the option to choose from the unscored, the once-scored, or the twice-scored composition boxes. The four-point third rater cutoff was determined by conference prior to the beginning of the term and was based on prior experience and results with this scale. Thus, by the end of the rating session all 143 papers had been read at least twice, 25 (about 17%) of the papers required a third reader, and the five norming papers had been read by all raters.

Input Dataset

For each subject the following data were keyed into the scores, and the subscores for the first, second, and--if necessary--third rater. The raw rater data included decimal values to one place because raters were permitted to assign " + .5" scores at any level. Next, the rater scores were averaged for each case (whether involving two raters or more). Later in the analysis the averaged data were rounded for the Rasch analysis portion of the study to facilitate use of the Microscale procedure.

Table 1 below presents descriptive statistics for the unrounded, averaged data less the five compositions used for norming purposes and less five subject records misplaced between the score-reporting and data entry stages. Thus the final data set analyzed and reported contained 133 cases.

Table 1

Descriptive Statistics: Unrounded Rater Averages
Fall 1985 ESL Composition Placement Test, UCLA.

Subskill	**Mean**	**S.D.**	**N**
Content	3.46	0.78	133
Organization	3.53	0.69	133
Expression	3.66	0.54	133
Structure	3.60	0.59	133
Mechanics	4.04	0.48	133

Table 2 below presents descriptive statistics for the rounded, averaged data.

Table 2

Descriptive Statistics: Rounded R ater Averages
Fall 1985 ESL Composition Placement Test, UCLA.

Subskill	**Mean**	**S.D.**	**N**
Content	3.57	0.83	133
Organization	3.70	0.74	133
Expression	3.79	0.62	133
Structure	3.75	0.63	133
Mechanics	4.10	0.55	133

As can be seen, rounding the data to achieve a five-point (non-decimal), averaged scale (suitable for the Rasch analysis reported below) had a minimal effect on the magnitudes of the means and standard deviations of composition ratings. Although the means and standard deviations increased slightly, the rank-ordering of the means remained unchanged.

Analysis Procedures

The analysis proceeded in two stages. First, correlation and multiple regression analyses were run on the unrounded data set to determine the contribution of each subscore to the total score for the purpose of suggesting appropriate subscale weights. Second, the rounded data were Rasch analyzed to address problems of person ability and category/continuum difficulty scaling and fit to the model.

The multiple regression analysis was done on SPSS Version H, using the "New Regression" procedure (Hull and Nie, 1981: 94-121). The solution employed the stepwise procedure. This routine was run on an IBM 3090 computer under the MVS operating system. Rasch analysis was done using Microscale Version 1.0 (Wright and Linacre, 1984). The Microscale analysis was done on an IBM PC (640K, two drives). This analysis generated logit scale values, fit statistic and standard error values (as explained in Wright and Stone, 1979) for all 133 persons, the five subscale categories, and the upper three of the five scale values. Because this data set was taken from the upper end of the ESL ability continuum in the UCLA ESL program, there were no low scale values of 1 and very few scale values of 2. Since Microscale does not analyze the bottom-most value of a scale (a feature somewhat analogous to the loss of a degree of freedom in inferential statistics), only the Rasch results for scale values 3, 4, and 5 are reported in this paper for each subscale.

Category

Mention needs to be made of the Rasch modeling assumption of unidimensionality. Unless this assumption is met with regard to the subscale categories, it would not be possible to position these categories along a single scoring continuum. The assumption would require that the hypothetical construct of composition writing performance be describable in a one-dimensional latent space incorporating the constructs of Content, Organization, Expression, Structure and Mechanics as defined by the scale. The usual method of testing this assumption involves use of principal components analysis to decompose the matrix of subscale category correlations into principal components. The finding of a primary, dominant factor accounting for at least 20 per cent of the matrix variance is taken as support for the presence of unidimensionality for purposes of application of item response theory (Reckase, 1979; Hattie, 1985).

In the present case this method is not applicable, since it is not considered a productive procedure to attempt to decompose a matrix of as few as five variables into other than a unidimensional solution (Cattell and Vogelman, 1977). This leaves alternative possible methods such as eyeballing of the correlation matrix (armchair factoring), consideration of estimates of internal consistency reliability, and examination of Item Response Theory (IRT) fit statistics. These procedures were followed in the present study. Although some correlational clustering was present, a high estimate of internal consistency (0.92) was obtained, no noticeable violation of category fit constraints appeared, and, thus, there was no convincing evidence that the rating data violated unidimensionality assumptions. Similarly, assumptions of non-speededness and local independence were not violated for the present data. A more rigorous test of unidimensionality might have been applied had there been additional subscale rating categories, say, 10 to 15 categories. Since all of the categories are language-related, and since they all involve the skill of writing, and because tests involving other, more diverse language skills have produced unidimensional solutions for students from this same population (Henning, Hudson and Turner, 1985) it is reasonable to anticipate a unidimensional solution.

Results

Observed Correlations

Correlational results are reported in Table 3 below. Note that there is a slight clustering tendency for subscales of Content and Organization and also for Structure and Expression. Note also that Content, Organization and Expression clearly bear the highest relation to total score, and that relation is similar in magnitude for each of these subscale categories.

Table 3

Subscore Correlation Matrix.

	C	O	E	S	M	T
Content	1.00					
Organization	.71	1.00				
Expression	.51	.39	1.00			
Structure	.38	.32	.73	1.00		
Mechanics	.22	.38	.36	.39	1.00	
Total	.81	.79	.79	.73	.58	1.00

Regression Analysis

The regression equation resolved on the final step had the standardized beta coefficients reported in Table 4 below.

Table 4

Standardized Regression Beta Coefficients and Suggested Weighting Factors.

Subscale Category	**Standardized Beta**	**Suggested Weighting**
Content	.337	1.261
Organization	.296	1.111
Expression	.235	.882
Structure	.256	.961
Mechanics	.208	.781
Total	1.332	4.996

The t-values associated with each of the standardized betas reported above was significant ($p<.05$). The suggested weights would be applied to observed scores in each subscale category to achieve maximal prediction of total score variance. Note that the rating total across all subscales was used as the best available dependent global estimate of writing ability. We are aware of possible deficiencies in this measure, and thus input of an overlap-corrected correlation matrix was also used in the regression procedure. Standardized betas are reported rather than raw score betas since the goal was not merely to predict or reproduce the total score, but to ascertain relative proportional contributions of subscales for derivation of weights.

Rasch Analysis

Microscale is unique in that, in addition to the typical person and item logit measures (cf. Rasch, 1960; Wright and Stone, 1979; Wright and Masters, 1982) it permits logit measurement for each of the five response scale values employed in the composition scale. Hence, the Wright and Masters (1982) refinement of the Rasch family of formulas can be applied to all three elements: persons, items, and response scale values. Potentially, analysis of response data can not only position these three elements along an ability/difficulty continuum, but can also identify persons, items and response categories for which the probabilistic predictions of the model are not met. When predictions of the model are not met, elements are said to misfit the model or exhibit a kind of response invalidity. Subsequent tables will report logit difficulty and fit statistics for subscores and rating scale values.

Table 5

Subscore Calibrations and Fit Statistics
UCLA ESL Composition Placement, Fall 1985.

Subscore	Logit Diff.	Error	Infit	Outfit
Expression	−.02	.14	−2.09	−1.19
Content	.57	.13	1.13	1.65
Organization	.23	.13	.22	.99
Structure	.08	.14	−.58	−.09
Mechanics	−.87	.14	−.74	.02

Table 6

Rating Scale Value Calibrations and Fit Statistics
UCLA ESL Composition Placement, Fall 1985.

Scale Value	Logit Diff.	Error	Infit	Outfit
3	−2.01	.17	15.00	8.39
4	−.85	.09	10.12	15.00
5	2.86	.11	1.74	−.62

Reliability Estimation

An added feature of the Microscale analysis is that reliability is calculated both as a global person separability index (0.92 for the overall composition grading scale) and as standard errors of measurement at each point along the scoring continuum. Note from Tables 5 and 6 that measurement error ranged from a low of .09 to a high of .17 logits. Both the global estimate of reliability and the individual standard error magnitudes were indicative of a high degree of measurement accuracy for this rating scale.

Person Misfit

Wright and Linacre (1984) discuss the difference between infit and outfit in the Microscale manual:

> We try to correct mismatches between items and people because they degrade the functioning of our measuring system. When item difficulty matches person ability (e.g., both are at 1.0 logits) we obtain the maximum amount of information about the variable and the person. The greater the difference between item difficulty and person ability, the less information the item provides. In order to minimize the influence of contacts between persons and items that are remote from one another, Microscale calculates an information weighted mean square residual, infit. The deviations used in calculating the outfit statistic are here weighted by the amount of information provided by each response . . . [cf. Wright and Masters, 1982: 99-100].
>
> In contrast to outfit, the infit statistic is not sensitive to outliers but focuses instead on the fit situation in the region where responses are delivering the most information [p. 4-18].

Careful scrutiny of the person fit statistics revealed five out of 133 persons exhibiting positive misfit, whether infit, outfit, or both. It was originally planned that these individuals be interviewed to determine patterns of behavior predictive of misfit. On subsequent consideration it became evident that person misfit could also derive as much from rater irregularities as from examinee irregularities. Therefore, all misfitting compositions were rerated independently of previous ratings. Rerating supported the consistency of previous ratings, and, therefore, attention was directed to the nature of the writing in the composition samples. The following observations were made:

1. The misfitting subjects represented a variety of language backgrounds (Spanish, Chinese and Iranian), so that the scale could not be said to be invalid for persons from any particular language background.

2. Misfitting subjects represented both sexes, so that there is no evidence of gender bias.

3. Each of the misfitting compositions was rated low on content with respect to the other subscale categories. Extreme problems with cohesion were evident in most cases. The misfitting performance resembled disordered language patterns as reported by Andreasen (1979). One of the students had been recommended for clinical counseling.

For such students, who may exhibit disordered thinking patterns, it may be the case that use of the composition grading scale would lead to invalid conclusions about student writing ability. No further patterns have as yet emerged from the analysis of misfit data.

Discussion of Subscores and Rating Scale Values

It can be seen from Tables 4, 5 and 6 that the primary operationalization of the composition variable is in the rating scale values. In other words, the rating values and not the subscale categories serve to bracket student ability. Indeed, without the facility of Microscale to analyze scale values, this might be overlooked. Table 4 indicates that the subscale scores differentially contribute to the total score. Table 5 shows that the range of logits for the subscale categories is quite narrow. Those logit values do not "bracket" the subjects. Rather, as seen in Table 6, the subjects are more nearly bracketed by the rating scale logits. This is an intuitively predictable finding, in keeping with previous work (Davidson and Henning, 1985). Generally, when one derives a scale such as this, there is a definite operationalization load on the rating scale value increments beyond that of the subscale categories.

This finding has implications for the use of a scale such as this particular composition rating instrument. It suggests that a great deal of training and norming effort should go into the clear understanding and interpretation of the meaning of each scale value increment; for example, what is the difference between a 4 and a 5 in Content?

This suggestion is upheld when one notes the much larger fit values in Table 6 than in Table 5. The rating scale increments seem to misfit at the mid-range values of the scale; i.e., 3 and 4. Wright and Linacre (1984) make the following observations about the fit statistics output by Microscale 1.0:

> No absolute rules are possible for the evaluation of fit. The judgments must be adjusted to the situation and to your aims. Generally we assume that items with outfits falling outside a band of 2.0 units on the outfit scale should be examined more closely [p. 4-13].
>
> As a general principle, whenever one of the fit statistics (outfit or infit) shows item fit values less than -2.0 we should look for factors that tie the items together more tightly than expected. Usually these factors will be influences on the measuring system which we want controlled. Sometimes the causes are a source of test bias, and should definitely be removed from the test [p. 4-17].
>
> Most psychometric procedures limit their attention to items which fit poorly because they contain too much noise (e.g., low item-by-test biserials would also identify [positive INFIT or OUTFIT items] as deviant). But items that fit too well because they contain too little noise, and hence signal an unexpected interdependence, also indicate unexpected irregularities in the measuring process. Overfit can often be explained by the presence of a second, unnoticed and usually unintended variable, positively correlated with the first, that, operating additively, creates a better fit than would occur if the items were more 'pure' in the measure of the intended main variable [pp. 4-15 to 4-17].

In the present analysis, the fit values for the subscale categories are all tolerable. The only marginally questionable fit statistic is the infit for the Expression subscore. Since it is negative, and bearing in mind the above comments, it may safely be assumed that the subscale categories fit model expectations well enough. However, the picture is somewhat more troubled with regard to the fit statistics for the scale values. They are intolerably high. This finding seems to suggest that the scale does not fit model predictions well along its scale value dimension, which further emphasizes the need for careful training and norming of raters concerning scale value differences.

One further comment can be made about the rating scale values. The sample under analysis was truncated in terms of the intended operation of this scale. The scale has been phased into the entire range of ESL courses at UCLA. At present, the only data that are available are from the upper courses, which are reported here. However, it is interesting to speculate on how the scale might operate at the lower end of the course spectrum, and indeed, such a study (a similar composition task across all levels) is in the planning stages.

Microscale 1.0 provides several interesting graphic displays through the use of its SuperCalc-3 spreadsheet capabilities (Sorcim, 1983). Figure 1 (see Appendix B) gives the relation of rating response probabilities to logit difficulty values. It is clear that the upper end of the scale, especially scale value 4, accommodates the majority of the response probability in this sample, as one might expect from a knowledge of the high general language proficiency represented.

What is more interesting is the comparatively low span of the latent difficulty continuum for which scale value 3 would be the most probable choice of the rater. This is shown in Figure 1 by the comparatively small unique, unshared probability area for scale value 3. This finding complements the finding of general misfit for scale value 3, and suggests that raters were not sufficiently uniform in their interpretation and application of the composition rating scale at that point. There may be a tendency for raters to identify extremes of performance with comparative ease and accuracy, but the mid-points of the scale encourage subjective judgments that are less reliable or valid.

Conclusion

In the introductory section of this study it was proposed that investigative results be reported for the UCLA ESL composition grading scale with regard to comparative difficulties of subscale categories, appropriate subscale combinatory weights, the nature of subscale-difficulty and performance-level intervals, global and local reliability estimates, person/category/level fit validity, and the nature of misfitting individual writing performance. This has been done in the present study. Tables 5 and 6 address the issue of comparative subscale difficulty (research question 1 above). Table 4 suggests subscale weightings (question 2). Figure 1 and its discussion address the probabilistic writing continuum in this dataset (question 3). Reported reliability values discussed above address scalar consistency (question 4). Tables 5 and 6, as well as the above discussions of fit address any misfit in subscale categories, scale values, or persons (question 5). Finally, the discussion of person misfit above identifies some interesting characteristics of student essays that should not be interpreted with the same decision confidence as the rest of the sample (question 6).

In general, results indicated that, while high measurement accuracy was observed, as reflected in global reliability estimation and local measurement error indices, misfit was present at the mid points of the performance scales. While it was readily possible for raters to identify highest performers, raters seemed to have difficulty distinguishing between performance ratings of three and four. This finding may suggest the need for a larger study to incorporate more learners at the lowest end of the ability continuum.

Misfitting subjects were not disproportionately representative of any one sex or language group background. Misfitting subjects did, however, exhibit interesting writing characteristics involving low content ratings relative to other subscale categories and demonstrating abnormally low cohesion, as would be representative of disordered language and thought patterns. This observation also merits follow-up and investigation with a larger sample of participants.

References

Andreasen, N. C. 1979: Thought, language, and communication disorders. Arch. gen. psychiatry, Vol. 36, Pp. 1315-1321.

Cattell, R. E. and Vogelman, S. 1977: A comprehensive trial of the scree and KG criteria for determining the number of factors. The Journal of Multivariate Behavioural Research. 12: 289-325.

Davidson, F. and Henning, G. 1985: A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing. Vol. 2, No. 2, Pp. 164-169.

Hattie, J. 1985: Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, Vol. 9, No. 2, Pp. 139-164.

Henning, G. 1983: Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. Language Learning. Vol. 33, No. 3, Pp. 315-332.

Henning, G. 1984: Advantages of latent trait measurement in language testing. Language Testing. Vol. 1, No. 2., Pp. 123-133.

Henning, G ., Hudson, T. and Turner, J. 1985: Item response theory and the assumption of unidimensionality for language tests. Language Testing. Vol. 2, No. 2, Pp. 141-154.

Hull, C. Hadlai and Norman H. Nie. 1981. SPSS update 7-9. New York: McGraw Hill.

Rasch, G. 1960: Probabilistic models for some intelligence and attainment tests. 1980 expanded edition. Chicago: The University of Chicago Press.

Reckase, M. D. 1979: Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics. Vol. 4, No. 3, Pp. 207-230.

Sorcim Corporation. 1983: SuperCalc-3 user's guide and reference manual: Documentation 1.0. Third edition. San Jose, Ca.: Sorcim Corporation.

Wright, B. and Linacre, J. M. 1984: Microscale manual for microscale version 1.2. Westport, Ct.: Mediax Interactive Technologies.

Wright, B. D. and Masters, G. N. 1982: Rating scale analysis. Chicago: MESA Press.

Wright, B. D. and Stone, M. H. 1979: Best test design. Chicago: MESA Press.

Appendix A

COMPOSITION GRADING SCALE

LANGUAGE

FUNCTION

Expression

5 PRECISE AND VARIED VOCABULARY/IDIOM USAGE; CONSISTENT AND APPROPRIATE VARIETY OF SENTENCE STRUCTURE; CONSISTENT AND APPROPRIATE REGISTER; CONCISE

4 GOOD AND VARIED VOCABULARY/IDIOM USAGE; A FEW MISUSED WORDS/EXPRESSIONS, APPROPRIATE VARIETY OF SENTENCE STRUCTURE, FLUENT EXPRESSION BUT INCONSISTENT OR INAPPROPRIATE REGISTER; FAIRLY CONCISE

3 USES BASIC VOCABULARY WELL; IDEAS NOT ALWAYS PHRASED IDIOMATICALLY; SOME VOCABULARY/IDIOMS MISUSED; ATTEMPT AT SENTENCE STRUCTURE VARIETY; LACKS AWARENESS OF REGISTER; MAY BE TOO WORDY

2 VOCABULARY TOO BASIC TO EXPRESS IDEAS EFFECTIVELY; VERY LITTLE SENTENCE STRUCTURE VARIETY; LITTLE AWARENESS OF REGISTER

1 INAPPROPRIATE USE OF VOCABULARY; NO SENTENCE VARIETY

WRITING

Organization

5 APPROPRIATE TITLE, EFFECTIVE INTRODUCTORY PARAGRAPH, THESIS MADE CLEAR WITH CLEAR CONTROLLING IDEA; EACH PARAGRAPH WELL FOCUSED AND WELL DEVELOPED; NECESSARY TRANSITIONS BETWEEN AND WITHIN PARAGRAPHS; CONCLUSION APPROPRIATELY TIES POINTS OF THE COMPOSITION TOGETHER

4. GOOD TITLE, INTRODUCTION, AND CONCLUSION; CEARLY ALL PARAGRAPHS HAVE A WELL-FOCUSED TOPIC SENTENCE, FULL SUPPORT, AND LOGICAL DEVELOPMENT; TRANSITIONS BETWEEN AND WITHIN PARAGRAPHS MAY BE MISSING OR INAPPROPRIATE

3. SOME PORTION OF THE COMPOSITION INCORRECTLY DEVELOPED; SOME ATTEMPT MADE TO STATE PURPOSE OF ESSAY, BUT CONTROLLING IDEA(S) UNCLEAR; PROBLEMS WITH ORDERING OF IDEAS IN BODY; TRANSITIONS BETWEEN AND WITHIN PARAGRAPHS MISSING OR INAPPROPRIATE; SOME EVIDENCE OF PLAN

2. MINIMAL OR INAPPROPRIATE INTRODUCTION OR CONCLUSION; PURPOSE OF ESSAY/CONTROLLING IDEA MISSING, POORLY STATED, OR UNCLEAR; IDEAS IN BODY ORDERED INAPPROPRIATELY; LITTLE EVIDENCE OF PLAN

1. LACK OF INTRODUCTION AND CONCLUSION; NO APPARENT PLAN

Content

5 GENERALIZATIONS ARE FULLY SUPPORTED AND MAKE SENSE; EXCEPTIONAL ANALYSIS OR INSIGHT SHOWN

4 GENERALIZATIONS ARE WELL SUPPORTED THROUGHOUT

3 NOT ENOUGH EXAMPLES OR SUPPORTING EVIDENCE FOR MAIN POINTS IN THE BODY; NEEDLESS REPETITION, MAJORITY OF MATERIAL ON TOPIC

2. LITTLE OR INAPPROPRIATE DEVELOPMENT OF THESIS; ENOUGH MATERIAL OFF TOPIC TO DISTRACT READER; PURPOSE OF ESSAY NOT CLEARLY IDENTIFIED

1. IDEAS NOT DEVELOPED, INSUFFICIENT SUPPORTING DETAILS, IDEAS TOO FEW

FORM

Structure

5 NATIVE LIKE USE OF VERB/MODAL FORMS, TENSE SEQUENCING, SUBJECT VERB AGREEMENT, ARTICLES, PREPOSITIONS, RELATIVE/ADVERBIAL CLAUSES, PHRASES; NO SENTENCE FRAGMENTS OR RUN-ONS

4 SOME GRAMMAR PROBLEMS WHICH DO NOT ADVERSELY AFFECT COMMUNICATION, e.g. VERB/MODAL FORMS, ARTICLES, PREPOSITIONS, PRONOUNS, RELATIVE/ADVERBIAL CLAUSES, PHRASES; ONE OR TWO SENTENCE FRAGMENTS OR RUN-ONS

3 IDEAS UNDERSTANDABLE DESPITE NUMEROUS, VARIED GRAMMAR PROBLEMS SUCH AS INCORRECT VERB TENSE, WORD FORMS, ARTICLE USAGE, WHICH HAVE A NEGATIVE EFFECT ON THE READER, MANY SENTENCE FRAGMENTS OR RUN-ONS

2 SOME EVIDENCE OF KNOWLEDGE OF GRAMMAR BUT SERIOUS GRAMMAR POROBLEMS INTERFERE WITH COMMUNICATION OF WRITER'S IDEAS

1 GRAMMAR PROBLEMS OBSCURE THE MESSAGE; MINIMAL CONTROL OF SENTENCE STRUCTURE

Mechanics

5 NEAT AND LEGIBLE WITH CORRECT USE OF MARGINS/INDENTATION, SPELLING PUNCTUATION, AND CAPITALIZATION

4 NEAT AND LEGIBLE WITH GENERALLY CORRECT USE OF MARGINS/INDENTATION SPELLING, PUNCTUATION, AND CAPITALIZATION

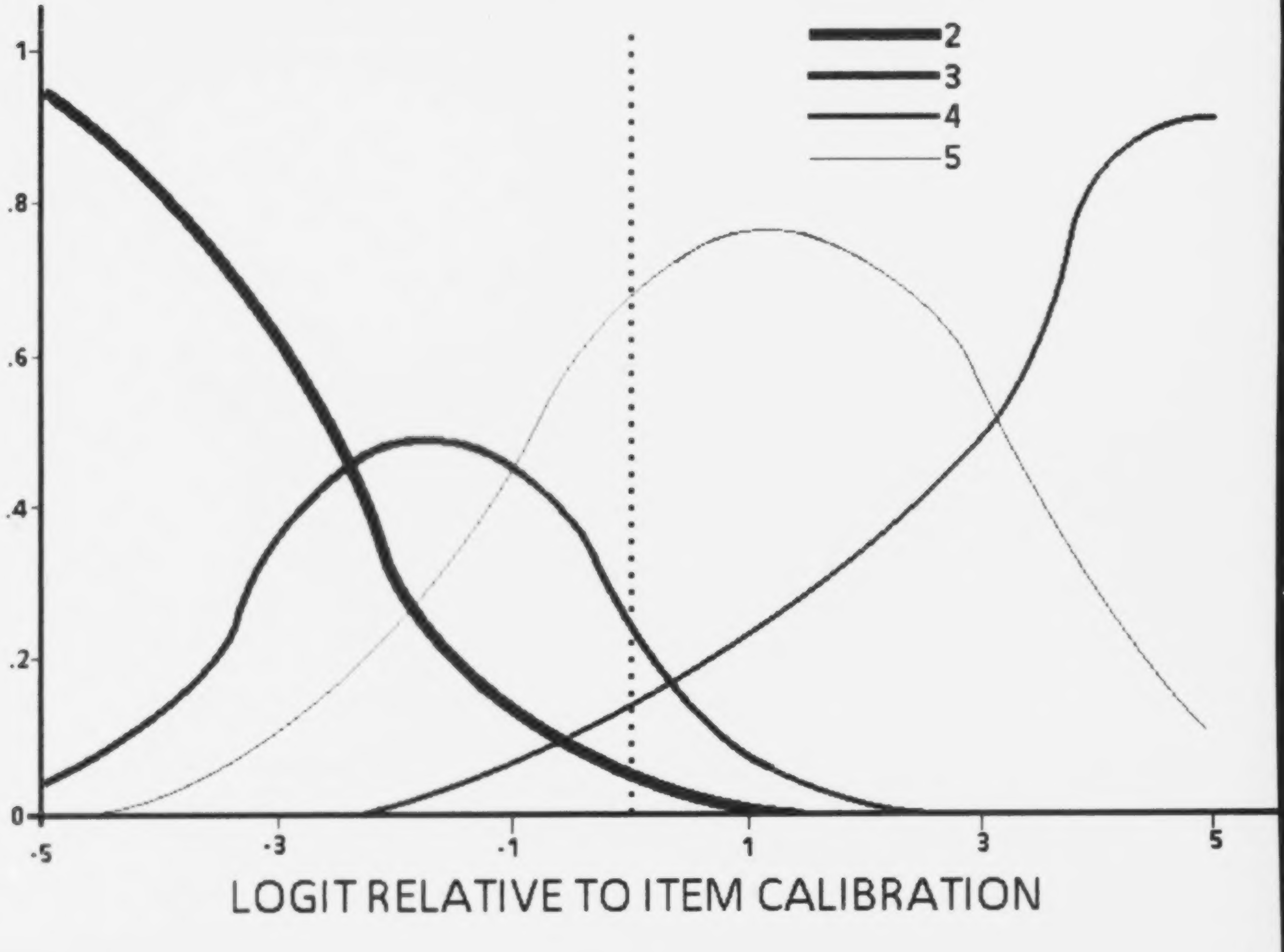
3. VARIOUS PROBLEMS WITH LEGIBILITY, MARGINS/INDENTATION, SPELLING, PUNCTUATION AND CAPITALIZATION

2 MAY BE DIFFICULT TO READ; FREQUENT PROBLEMS WITH MARGINS/INDENTATION SPELLING, PUNCTUATION, AND CAPITALIZATION

1 MAY BE ILLEGIBLE; SERIOUS PROBLEMS WITH MARGINS/INDENTATION, SPELLING PUNCTUATION, AND CAPITALIZATON

FIGURE 1

COMPOSITION RATING SCALE - F85 PLACEMENT
PROBABILITY OF RESPONSE BY CATEGORY



THE GRAPHIC REPRESENTATION OF LANGUAGE COMPETENCE:

MAPPING EFL PROFICIENCY USING A MULTIDIMENSIONAL SCALING TECHNIQUE

Russanne Hozayin
American University in Cairo
English Language Institute

In a recent review of the literature on assessing the dimensionality of tests and items, Hattie (1985) pointed out that, "one of the most critical and basic assumptions of measurement theory is that a set of items forming an instrument all measure just one thing in common" (p. 139). He further noted that the assumption of unidimensionality underlies most measurement models. It also plays a crucial role in the interpretation of test results.

Two areas which have vital links with the dimensions underlying a set of items are **construct validity** and **latent structure analysis**. Traditionally, dimensionality has been closely related to the validity of the test, especially to construct validity, since investigators have often tried to assess the goodness-of-fit of a set of data to a theory-based model. That is, one approach to answering the question, "What is the test actually testing?", is to ask, "How many significant, identifiably separate factors is the test testing?"

In language ability testing, determination of underlying structure has virtually always taken the form of factor analysis of an intercorrelational matrix of subtest and/or total test scores. These subtests/test scores have most often represented a language skill (vocabulary); a specific format (cloze); or a specific test (TOEFL). (See Part I in both Oller & Perkins [1980] and Oller [1983] for detailed discussions of the pros and cons of these procedures.)

Besides construct validity, another area linked to dimensionality is that of data analysis models. In recent years, there has been a dramatic increase in the use of latent structure models[1] to analyze a wide variety of data (cf. Bergan & Stone, 1985; Young & Tanner, 1984).

Virtually all measurement models (as well as theory-based models) assume unidimensionality of test items. This assumption is particularly important for latent trait/item response theory models, like Rasch, which are

---

The author would like to gratefully acknowledge the assistance of the following people in the preparation and presentation of this paper: the faculty, staff and students of the English Language Section of the Department of Public Service, American University in Cairo; the staff of the Computer Center of AUC; and the Research and Conference Grant Program of AUC.

[1] Latent structure models are, according to Young & Tanner (1984), of four main types: latent trait, latent class, mixture analysis, and factor analysis. For one cogent approach to classification of structural models, see their article.

becoming more and more widely used in the analysis of language test results (Henning, 1984).

Therefore, in terms of the determination of construct validity of the test items as well as the fit of the measurement model to the data, the question of dimensionality is one of vital importance, both theoretically and practically. However, as Hattie (1985) has also pointed out, "despite its importance, there is not an accepted and effective index of the unidimensionality of a set of items" (p.139).

One major reason for the lack of progress in determining dimensionality of tests is that of confusion over definition. Although a clear, acceptable definition of unidimensionality is not yet forthcoming, certain clarifications have been made recently. For example, Hattie (1985) has drawn distinctions between "homogeneity," "internal consistency" and "reliability" on the one hand and "unidimensionality" on the other.

In his extensive review of more than 90 articles, Hattie made the very important point that high reliability (i.e., alpha and its derivatives) by itself may not be an indication of unidimensionality. This conclusion is crucial in view of the importance of reliability as an index of unidimensionality--"Perhaps the most widely used index of unidimensionality has been coefficient alpha (or KR-20)" (Hattie, 1985, p.143).

Henning, Hudson and Turner (1985) provided further insight when they noted that "no test can strictly be said to measure one and only one trait" (p.142). Thus, a test will be unidimensional when it "measures a primary or dominant factor" (Henning, et al., 1985, p.145) or when "one latent trait underlies the data" (Hattie, 1985, p.139). In effect, the question has been rephrased from the more simplistic "What is unidimensionality?" to the more complex, yet potentially more realistic, "How close is a set of items to being a unidimensional set?" (Hattie, 1985 p.159) and "To what extent must [unidimensionality] be present in order for latent trait models to apply?" (c.f., Henning, et al., 1985, p.142).

Thus, attempts at clarifying definitions have led to some refinements of one of the questions to which answers are being sought. However, beyond problems with definitions there are difficulties with the statistical techniques which have been applied to the problem of unidimensionality, particularly with factor analysis.

In the past, there have been two main approaches to the assessment of dimensionality of tests or items:

(1) The investigator has a set of data which he or she wishes to analyze in order to reveal the number of factors being measured.

data ---> via measurement method ---> model

(2) The investigator has posited a model and he or she wishes to assess the goodness-of-fit of that model to a set of data.

model ---> via data analysis ---> fit

For both of these approaches, in language testing, subtest scores or scores from a variety of instruments have been correlated, and the resulting matrix has been factor analyzed--by using either traditional factor analytic methods (Oller & Hinofotis, 1980) or confirmatory factor analysis (Vollmer, 1985). In order to be able to assess the relative success of these past approaches, both the growing importance of item level analysis and the increasing criticism of factor analysis on both technical and substantive grounds must be taken into consideration.

As has already been noted, the increasing interest in item-level analysis in language testing has been due in large part to the increased use of latent trait analyses, which have been stimulated by the rapid developments in computer technology as well as the general spread of cognitivism. In addition, with such developments as item banking and computer adaptive testing, wherein relatively few items can give the same level of information on the subject's ability as can a longer paper-and-pencil test, item-level analysis is becoming more and more important. Besides the importance of item-level analysis, there are two critical points about factor analysis that are pertinent to the current study, including assumptions about the linearity of relationships among the factors as well as requirements for certain types or levels of data when using different latent structure techniques.

First, factor analysis--as a latent structure technique--is intended to reveal the "hidden" relationship among the variables and thereby lay bare the underlying structure of the trait (items) being measured. However, in nearly all previous research studies which used factor analysis, it was assumed that the relationships between the scores used to generate the correlation matrix were linear. This assumption must be looked at more closely, in the light of recent developments in measurement models and the assumptions underlying them.

As Vollmer & Sang (1983) pointed out,

> . . . possibly a linear outlook on language proficiency is far too simple...it is important to stress that . . . differences in the interaction between different levels and aspects of ability on the one hand and between object (items) and subject (person) on the other hand can neither be discovered nor be described by means of classic factor analysis (p.72).

Other writers concur, as when Hattie (1985) stated that, "a further problem of many procedures is that a linear model cannot be assumed. When items are scored dichotomously, then the use of a linear factor model [is] not appropriate since [it] assumes linearly related variables" (p. 158).

It must be recalled here that there are in fact at least two relationships whose linearity/non-linearity must be accounted for. Diagram One represents the relationship between the manifest scores (or item responses), on the one hand, and between the surface representative and the latent trait which the test (item) score is assumed to reflect, on the other. The top line depicts the manifest relationship between two subtest/total test scores (traditionally calculated by a Pearson product moment correlation coefficient); the second line, the manifest relationship between two 0/1 items (traditionally calculated by a phi or tetrachoric coefficient).

Diagram One

The Manifest and Latent Relationships Between Subtest/Total Test Scores or 0/1 Items

subtest/test--------[surface/manifest]--------subtest/test

0/1 item--------[surface/manifest]--------0/1 item

linear?

deep (latent trait)

The vertical line in Diagram One represents the relationship between the manifest score, whether subtest/total test or 0/1 item, and the posited latent trait. This relationship is based on a theoretical model and its nature at present must be considered to be largely speculative, due to the paucity of results based on analysis of actual data sets.

The second point related to the use of factor analysis is that of the level of the data. One of the crucial factors in the choice of a statistical technique for a given data analysis situation is whether the data is categorical, rank-order, continuous or ratio. Young and Tanner (1984) have pointed out that factor analysis is appropriate when both the manifest and latent variables are continuous, while latent trait techniques are suitable when the manifest variable is dichotomous and the latent is continuous. Therefore, at the item level, if the items are dichotomous, factor analysis is inappropriate. It may be concluded, then, that if item level test analysis is to be undertaken, then linear factor analysis is in fact an ineligible analytic technique, both due to the questions raised above concerning the nature of the manifest-latent relationships and to the level of the data involved in 0/1 item analysis.

As an alternative to linear factor analysis--which was formulated in a time when technology was relatively embryonic and whose assumptions may well be unrealistic in terms of reflecting the nature of the relationships among the structures underlying person responses, it was decided to try another approach to assessing unidimensionality--that of non-metric multidimensional scaling, using item-level analysis.

Multidimensional Scaling

Multidimensional scaling (MDS) is a statistical technique often compared to factor analysis (FA):

> they are both used to study the structure of objects; they require similar input data, proximity measurements defined over pairs of objects; [and] they both represent structure in terms of spatial coordinates (Davison, 1985, p. 94).

As Schiffman, Reynolds and Young (1981) point out, however, there are certain crucial differences between MDS and factor analysis, particularly that "the MDS model is based on distances between points whereas the FA model is based on angles between vectors...[the former being] easier to interpret" (p. 13). In addition, they point out that "FA often results in a relatively large number of dimensions mainly because most procedures are based on the assumption of linear relationships between the variables...[MDS] does not contain this assumption" (p. 13).

An important advantage of MDS over factor analysis is the graphic representation of the data which comprises part of the MDS output. To obtain this pictorial analysis, MDS programs present numerical representatives of objects which are similar to one another (in this study, correlations between 0/1 responses to items) as points close to each other in a spatial map. Objects which are dissimilar are represented as points distant from one another (Schiffman, et al., 1981). Since MDS uses distance instead of angular separation (which is used by factor analysis), there is no assumption of linearity required (Coxon, 1982).

Subkoviak (1975), in reviewing the use of MDS in educational research, confirmed that MDS "can be used to discover the underlying structure of a test" (p. 412) at the item level, although it has rarely been used for this purpose. More recently, Davison (1985) noted that, "Factor analysis... remains the most common method used to analyze the structure of tests or test items. More and more, however, MDS has been recommended for the same purpose" (p. 94). Those who have used it at the item level include Farley & Cohen (1974), Karni & Levin (1972), and Napior (1972).

Prior to this study, as far as can be determined, MDS has never been used to investigate the structure underlying an ability test at the item level as an alternative to factor analysis. After reviewing the literature, though, the present investigator concluded that because MDS does not require an assumption of linearity, and because its output is in the form of graphical representations of the data, it would be a viable alternative to factor analysis.

However, it must be assumed for the purposes of this study that the relationship between answers to each pair of items for a given sample of subjects can be analyzed to reveal information about the structure of the language competence of those subjects. With this assumption in mind, the methods used in the study, along with the results and conclusions, will now be presented.

Method

Subjects

The subjects who participated in this study were part of a group of more than 8500 students who take adult English courses at the Department of Public Service (DPS), American University in Cairo, each term. These courses are non-intensive (3 hours per week), and last for 12 weeks. In the majority of cases, students do not take these courses in order to satisfy an immediate job or educational requirement, but in order to improve their English skills, possibly to obtain a better position in the future.

The subjects in the current study were 149 adult Egyptian basic-level EFL learners, aged between 18 and 55. These people were placed in Level 3 (out of 17 levels) on the basis of: (1) their scores on a standardized test written and used by the DPS for placement, or (2) their having passed Level 2. Following their placement at Level 3, they were then assigned at random to one of the 36 Level 3 classes which were offered during the Fall, 1985 term. Students in all 36 classes took the post-test which formed the basis of this study. Class-size averaged between 14 and 18 students.

Instrument

The test used in this study was one of three forms taken by the Level 3 students as a post-test. These three forms were the result of a pre-testing process which took place over a 9 month period and included approximately 2500 students from Levels 2 through 5 of DPS. Each form contained 36 questions, including six multiple-choice grammar and vocabulary items, two 10-item cloze passages and one 10-item elision passage. In the latter, the students were required to delete 10 extraneous words from the passage. The ten target words, however, were not indicated in any way, so the students had a discourse level rather than a discrete-point type of task (although, due to the low level of their English ability, they were assisted in completing the task by being told the precise number of extra words which occurred in the passage).

All items were based on the Key to English textbooks used in the DPS courses. These books were designed and written by the DPS curriculum developers specifically for Egyptian students. Prior to the item writing process, a detailed content analysis of the textbooks was carried out (Hozayin, 1986), with the aim being to establish a basis for determining the content validity of the test.

After each item pre-testing session, the results were item-analyzed, using the Rasch BICAL program (Wright & Stone, 1979). Items for subsequent forms were selected on the basis of their total T-Fit (a BICAL statistic) being equal to or greater than one standard deviation below the mean total T-Fit. Other criteria that were taken into consideration in selecting items were the discriminability of the item (.80 or higher) and the point biserial correlation (.25 or higher). These additional criteria were used since at times the T-Fit of the item may be acceptable, but the item may not add much information concerning the students' ability, as, for example, when an item is too easy.

This pre-testing process resulted in short tests (36 to 38 items) with relatively high KR-20 reliabilities (typically, between .86 and .92 for the total test). As may be seen in Table 1, the KR-20 reliability of the total test was .89, while those of the two cloze and the one elision passage were .79, .71 and .80, respectively.

Table 1 also shows the means and standard deviations of the total test and the three passages. The mean of 18.61 (out of a possible 36), which is lower than one might expect on a post-test, probably resulted from a combination of circumstances. First, the item pre-testing process showed that items which fit the students prior to the course were not suitable post-course; thus, the pre-and post-tests were not exactly the same, although they did share some items. Second, the non-intensive nature of the program and the low level of the students, as well as their lack of immediate, specific and vital purpose in taking the course, could be acting (and inter-acting) to depress the mean.

Another factor, mirrored in the relatively high standard deviation (7.25 for the total test), is that although the students were all basic-level, they were still characterized by a range of abilities with their general level. This is because placement, as mentioned above, may have been by screening test scores or by the students having passed the previous level. This two-pronged system, when combined with student and program attributes, led to diversity of ability among the students. It also may be in part responsible for the elevated reliability of the test.

The subtest and total test intercorrelation coefficients contained in Table 1 show that there were moderate correlations between the three subtests (from r = .4294 to r = .5619). It should be mentioned here that multiple choice items 5 and 6 were included in the correlation matrix because they were the only multiple choice items identified by the multidimensional scaling analysis (see below) as being significantly related to the trait under discussion. In addition, since the six multiple choice items were analyzed as individual items, not as a subtest, means and standard deviations were not reported for these items by themselves or as a subtest.

Table 1

Test Statistics

Test Subtest	**Mean**	**SD**	**DR20***	**No. of Items**
Total	18.61	7.25	.89	36
Cloze 7-16	4.76	2.72	.79 (.88)	10
Elision 17-26	5.80	2.43	.80 (.88)	10
Cloze 27-3	4.83	2.30	.70 (.83)	10

*Due to the small number of items per subtest, the Spearman-Brown formula was applied. The estimated r for 20 items is given in parentheses.

N = 149 for all tables and graphs.

Table 1 (Continued)

PMC Intercorrelation Matrix of Items, Subtests and Total

Item 5	1.000					
Item 6	.2617	1.000				
Cloze 1	.4813	.4407	1.000			
Elision	.3610	.2810	.5389	1.000		
Cloze 2	.3340	.3370	.5619	.4294	1.000	
Total*	.5002	.4381	.7156	.5585	.5702	1.000
	Item 5	**Item 6**	**Cloze 1**	**Elision**	**Cloze 2**	**Total**

*These correlations have been corrected for part-whole overlap.

Data Collection Procedures

The tests were administered by the classroom teachers in the last week of class. The three forms were distributed at random to the students. Afterwards, the tests were returned to the investigator who scored them according to criteria derived from the responses of twenty EFL teachers (12 native speakers of English and 8 non-native speakers) to the cloze and elision passages. The responses of the native speakers were used in the cloze correction process. In the rare cases when a potentially correct response was not included among the answer given by the native speaker teachers, two native speakers (the investigator and one teacher) decided whether to score the response as being correct or not.

The elision passage was scored by giving credit only when the ten items of interest were crossed out. If other words in the passage were crossed out, no points were deducted. Rarely did a student cross out more than ten words. If they got a low score on elision it was usually because they couldn't perceive that a word was extraneous (thus deleting fewer than 10 words), or they perceived the extra word incorrectly (deleting 10 words, but not the correct ones). Scoring procedures for elision are obviously more complex than those used with cloze. The scoring procedure applied in this study was used in order to obtain a stable data set for inclusion in the analysis. However, in the future, alternative procedures could be applied and the resulting data could be re-analyzed following the steps outlined below.

Data Analysis

The scores for the 36 items were first analyzed using BICAL. Then a Pearson inter-item correlation matrix was generated. These correlation coefficients were then analyzed by KYST, a non-metric MDS program (see Kruskal, Young,

& Seery, 1973, for a complete description of KYST). Since a non-metric MDS application was being used, Pearson coefficients[2] were acceptable input for the KYST program (cf. Coxon, 1982, for a discussion of the types of data which may be used in MDS programs).

The KYST output, which included data plots for 1 through 5 dimensions, reached a minimum configuration in all five dimensions, with the following stress rates (error or noise)[3]:

1 dim = .418; 2 dim = .267; 3 dim = .198;

4 dim = .155; 5 dim = .126

Although there was a noticeable decline in stress as the number of dimensions increased, the level of stress in five dimensions was still relatively high, according to guidelines offered by Kruskal and Wish (1978). This may have been due to: (a) the basic level of the students; (b) the fact that the analysis was at the item level; (c) the nature or number of the items; or, most likely, (d) a combination of these factors. The implications of these factors in terms of stress are discussed below.

Results and Discussion

Since correlations were used in the KYST analysis, the emphasis was on small distances (large similarities) between pairs of items. Therefore, the results are presented in the form of neighborhood analyses (cf. Kruskal & Wish, 1978). That is, as shown in Table 4, below lines were drawn between the 32 item pairs with the highest correlation coefficients, ranging from .384 through .581. These coefficients represent 5 percent of the coefficients used in the KYST analysis, which totalled 630, this being the total number of coefficients arising from correlating 36 items and creating a half-matrix (i.e., number of items divided by number of items minus 1, divided by 2 to give a half-matrix).

Appendix A contains the 22 items identified by this procedure, while the item statistics, including Rasch total T-fit, difficulty and point-biserial correlations are given in Table 2. In addition, the item characteristic curve data are given in Table 3, showing the percentage of subjects who got each item correct, for six ability groupings. These 22 items may be said to represent the core of the learner's EFL competence, as measured by this test.

---

[2] Although such a discussion is beyond the scope of this paper, there are several very important considerations in selecting correlation coefficients for use in multivariate procedures. The interested reader should consult Coxon (1982) for specific information on MDS, Harris (1975), and Thorndike (1978).

[3] While the definition of stress is fairly standard in any work on MDS, for example, "[stress is] a measure that shows how far the data depart from the model" (Schiffman, et al., 1981), the interpretation of the stress coefficient and determination of what an acceptable level of stress is in a given case are both rather vexed subjects in the literature. Kruskal and Wish (1978) offer the most fundamental explanation, which a newcomer to MDS may wish to pursue.

Table 4, (a, b, c, d) which offers a graphic portrayal of the structural relationship of the items that make up this core, contains a neighborhood KYST analysis of 22 core post-test items for these 149 subjects, for 2 by 1 dimensions in 2-, 3-, 4- and 5-dimensional space. In order to assess the dimensionality of the data, these diagrams were first examined for patterns in the data according to four types of criteria: (1) format/item type; (2) item fit (Rasch T-Fit from BICAL); (3) item content; and (4) item difficulty.

Although not conclusive, the only clear pattern was based on clustering by item type, in that several of the elision items were grouped together on the left of the diagram, while most of the cloze items were grouped in the center (see Table 4). Thus, the cloze items appear to form the core of the diagrams, being the items which are most representative of the total test score (cf. their higher point-biserial correlations). Further, it is interesting to note that there are two types of relationships possible among the items--spatial proximity and correlational proximity. For example, the four cloze items from one of the two passages included in the test (i.e., items 33-36) formed one separate cluster (in terms of their high intercorrelations). At the same time, they were also in fairly close proximity to other cloze items, certainly closer to them than to the elision items which were on the opposite side of the diagram.

These relatively distinct groupings of elision and cloze items may indicate that two different cognitive activities are being measured by this test. It should also be kept in mind that while the students were familiar with cloze procedures, they were not as familiar with elision procedures, according to several of the teachers. This lack of familiarity may be responsible for the apparent division between these two formats. In subsequent test forms, instructions were given in Arabic and teachers reported that difficulties associated with elision were much reduced. The results of these tests are being examined to determine if patterns similar to those found in this data set are typical or if these patterns are related to the subjects' unfamiliarity with the elision format.

A second approach to evaluation of the diagrams was to examine the 2 by 1 dimensional structure underlying the 22 core items. When each of the four dimensions (Table 4) was assessed, it was noted that the five-dimensional portrayal of the relationship between dimensions one and two was the most parsimonious, since the items which were most closely related (i.e., which had the highest correlations) were also closest together in space in the five dimensional diagram. This is a further indication that these items are measuring multiple objects.

What conclusions may be drawn from this application of multidimensional scaling to an item-level EFL ability data set? With regard to the issue of unidimensionality, in terms of both construct validity and models for latent structure analysis, several questions may be raised, chief among them being whether it is possible that items might be multidimensional (as they appear to be here), while the total test could be considered to be unidimensional in the currently accepted definition of the term. However, even if this were found to be true, it is clear that if item level analysis is to be pursued, even the recent, more explicative definitions of unidimensionality mentioned in the Introduction will need to be reconsidered.

Equally clearly, much more use needs to be made of MDS at the item level. It should be mentioned here that this application is exploratory in nature rather than being confirmatory MDS (cf. Traub & Lam, 1985; Young, 1984). In future studies, if insights obtained into item-level analysis of ability by similar applications of exploratory MDS warrant, then confirmatory MDS might be used.

Another consideration is that this particular set of data has a relatively high amount of stress. Possible explanations already offered are related to the ability level of the students, the item-level of the analysis, the number and type of items or a combination of these factors. It might be noted in this connection that a reanalysis by the present author of the Oller-Hinofotis (1980) [following Pang's (1984) example] total test/subtest data using KYST revealed a much lower stress value (.01 in five dimensions) than was found for the current item-level data. Thus, it may be that item-level data have more noise, or that this particular data set has more noise.

Further analyses at the item level are needed in order to shed more light on the problem of stress or noise in the data. Future studies could focus on item analysis of longer tests as well as of standardized tests, using results obtained from EFL students of differing abilities.

It is readily apparent that that MDS is a complex procedure. However, the current investigator believes that this complexity is necessary in order to be able to reflect the intricacies of the cognitive processes which, it is assumed, item-level analyses of ability test results can reveal. Somewhat offsetting the possible disadvantages of this complexity is the spatial output one may obtain, which may be analyzed to reveal information concerning the product of language learning and--after much more use is made of it--possibly the process.

The major drawback to multidimensional scaling at this point is our lack of ability to interpret the multidimensional findings more meaningfully, especially in terms of the implications they may have for the construct validity of language ability tests, further explication of which was one of the purposes stated at the outset of this paper. Capitalizing on the spatial nature of the MDS output is a potentially fruitful path to meaningful interpretation of findings, since several other disciplines (such as sociology, cognitive psychology, ecology, biology, and physics) have established guidelines for analysis of spatial data. The current investigator is pursuing several analytic methods used in other fields in an attempt to shed light on the findings reported herein.

With regard to the assumption of unidimensionality underlying latent trait or latent structure analyses, it may be noted that investigations into dimensionality as they now stand include a large element of informed intuition and judgement on the part of the investigator, which may also be characterized as subjectivity. A certain amount of this seems to be unavoidable given our present state of knowledge (i.e., the lack thereof).

When a complex subject is being explored, summary methods of analysis -- although perhaps more parsimonious -- may not produce results which can adequately reflect the nature of the object of study. In order for investigators to begin to draw nearer to the desiderata of parsimony and elegance in theory building, model testing, and data analysis, approaches which are systematic yet which produce richer, more explicative and potentially more meaningful results are necessary. It is hoped that the application of MDS to item level ability data described in this paper will prove to be a step in this direction.

References

Bergan, J. & Stone, C. 1985. Latent class models for knowledge domains. *Psychological Bulletin*, 98(1):166-184.

Coxon, A. 1982. *The User's Guide to Multidimensional Scaling*. New York: Heinemann Educational Books.

Davison, M. 1985. Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin*, 97: 94-105.

Farley, F. & Cohen, A. 1974. Common-item effects and the smallest space analysis of structure. *Psychological Bulletin*, 81: 766-772.

Harris, R. 1975. *A Primer of Multivariate Statistics.* New York: Academic Press.

Hattie, J. 1985. Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9 (2):139-164.

Henning, G. 1984. Advantages of latent trait measurement in language testing. *Language Testing*, 1:123-133.

Henning, G., Hudson, T. & Turner, J. 1985. Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 2:141-154.

Hozayin, R. 1986. Using curricular domain assessment to establish the content validity of an EFL achievement test. A paper presented at the 6th National Symposium on English Teaching in Egypt: Testing and Evaluation, Ismalia, Egypt, March 25, 1986.

Karni, E. & Levin, J. 1972. The use of smallest space analysis in studying scale structure: An application to the California Psychological Inventory. *Journal of Applied Psychology*, 56:341-346.

Kruskal, J.B. & Wish, M. 1978. *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.

Kruskal, J.B., Young, F.W. & Seery, J.B.1973. How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Murray Hill, NJ: Bell Labs., mimeo.

Napior, D. 1972. Nonmetric multidimensional techniques for summated ratings. In R. Shepard, et al., *Multidimensional Scaling*, Vol. 1. (New York: Seminar Press).

Oller, J., ed. 1983. *Issues in Language Testing Research*, Rowley, MA: Newbury House .

Oller, J. & Hinofotis, F. 1980. Two mutually exclusive hypotheses about second language ability: indivisible and partially divisible competence, in Oller & Perkins, 1980.

Oller, J. & Perkins, K. 1980. Research in Language Testing, Rowley, MA: Newbury House.

Pang, L.Y. 1984. Is there a global factor of language proficiency? A critique of Oller and Hinofotis (1980). IRAL, 22:203-212

Schiffman, S., Reynolds, M. & Young, F. 1981. Introduction to Multidimensional Scaling, New York: Academic Press.

Subkoviak, M. 1975. The use of multidimensional scaling in educational research. Review of Educational Research, 45:387-423.

Thorndike, R. 1978. Correlational Procedures for Research, New York: Gardner Press (distributed by Halstead Press).

Traub, R.E. & Lam, Y.R. 1985. Latent structure and item sampling models for testing. Annual Review of Psychology, 36:19-48.

Vollmer, H. 1985. Models of second language competence: A structural equation approach. A paper presented at the 7th Annual Language Testing Research Colloquium, Princeton, N.J., April 6-9, 1985.

Vollmer, H. & Sang, F. 1983. Competing hypotheses about second language ability: A plea for caution, in J. Oller (1983), pp. 29-79.

Wright, B. & Stone, M. 1979. Best Test Design, Chicago: Mesa Press.

Young, F.W. 1984. Scaling. Annual Review of Psychology, 35:55-81.

Young, M.A. & Tanner, M.A. 1984. Recent advances in the analysis of qualititative data with applications to diagnostic classification. In Gibbons, R.D. & Dysken, M., eds. Statistical and Methodological Advances in Psychiatric Research. New York: Spectrum.

Appendix A

Items Included in Graphs

Items (M - Multiple choice vocabulary; C - Cloze; E - Elision)

M5 - I can't hear you because the boys are very __________.
1. noisy 2. busy 3. tired 4. important

M6 - Please find this word in the __________ .
1. calculator 2. dictionary 3. typewriter 4. adding machine

C7-16 (Items 8 and 15 not included in graphs):

Mona has a big family. __C7__ has three brothers and three __C8__. Two of her brothers __C9__ engineers and the other is a student. __C10__ of her sisters is married, but the __C11__ are studying business at the university. Mona's __C12__ is an accountant and __C13__ mother is a teacher __C14__ a girl's school. Because __C15__ are many people in the family, __C16__ must work hard to succeed.

E17-26 (Items 25 and 26 not included in graphs):

"Hello, John. How are you?" "Fine, thanks, Ahmed. How are you?"
"Fine, thanks. I'm going [E17 at] to the Pyramids. Can you come with [E18 to] me?"
"No, I'm sorry. I [E19 am] can't. I'm busy?"
"That's too bad. Next week I'm going [E20 go] to visit my sister in [E21 the] Ismalia. Can you go with [E22 my] me?"
"OK. What time are you going [E23 at] to leave?"
"I am leaving [E24 in] at 8 a.m. Where can I pick [E25 get] you up?"
"In front of my house [E26 family]."
"OK. See you next week, John."

C27-36 (Items 27 through 32 not included in graphs):

"What's the matter, Nabil? Are __C27__ tired?"
"No, I'm not. I'm __C28__. Is there a restaurant __C29__ here?"
"Yes, there's a good one __C30__ there."
"What's the food __C31__?"
"It's very good. What do you __C32__ to eat?"
"I'm not sure. __C33__ there any kofta?"
"Yes, there's __C34__, and there's kebab too."
"Are __C35__ any good salads?"
"Yes, there __C36__ some."
"OK. Let's go eat."

Table 2

Rasch Statistics

Item	T-Fit	Diffic	pbs
M5	-.70	.06	.55
M6	-.28	.06	.51
C7	-.58	-1.42	.43
C9	-.83	.49	.55
C10	-1.56	.26	.63
C11	-1.51	1.36	.59
C12	-1.93	.51	.66
C13	-1.47	-.39	.61
C14	-1.04	-.24	.57
C16	-1.27	.31	.60
E17	-.50	-1.18	.44
E18	-1.29	-1.03	.55
E19	-.68	-.94	.50
E20	-.27	-.56	.48
E21	-.43	.96	.48
E22	-.79	-1.42	.48
E23	-.91	-1.18	.48
E24	-1.06	-.76	.55
C33	-.94	-.37	.56
C34	-.85	-.42	.54
C35	-.69	.03	.55
C36	.12	-.09	.47

T-Fit = Rasch item total t-fit Diffic = Rasch item difficulty
pbs = point biserial correlations

Table 3

Item Characteristic Curve

Item	Lowest	2	3	4	5	Highest
M5	.16	.19	.50	.65	.77	1.00
M6	.28	.31	.39	.54	.92	1.00
C7	.60	.73	.89	.96	1.00	1.00
C9	.08	.12	.32	.46	.69	.92
C10	.00	.23	.32	.65	.77	1.00
C11	.00	.00	.04	.02	.58	.77
C12	.08	.00	.29	.38	.88	.92
C13	.12	.50	.57	.85	.96	.92
C14	.16	.31	.57	.88	.81	1.00
C16	.08	.15	.29	.62	.81	.92
E17	.40	.73	.93	.92	1.00	.92
E18	.32	.62	.93	.96	.96	1.00

Table 3

Item Characteristic Curve (Continued)

Item	Lowest	2	3	4	5	Highest
E19	.36	.65	.75	.92	1.00	1.00
E20	.32	.46	.68	.85	.92	.92
E21	.04	.08	.14	.38	.50	.96
E22	.52	.73	.93	1.00	1.00	1.00
E23	.36	.81	.89	.92	.96	1.00
E24	.32	.54	.71	.96	.92	1.00
C33	.20	.35	.64	.88	.85	1.00
C34	.20	.46	.61	.92	.81	1.00
C35	.16	.31	.39	.58	.88	1.00
C36	.28	.42	.32	.62	.88	1.00
Score range:	1-10	11-14	15-19	20-24	25-28	29-35
No. of SS:	25	26	28	26	26	13

Note: Figures represent percentages in each ability grouping

Table 4a

TWO DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

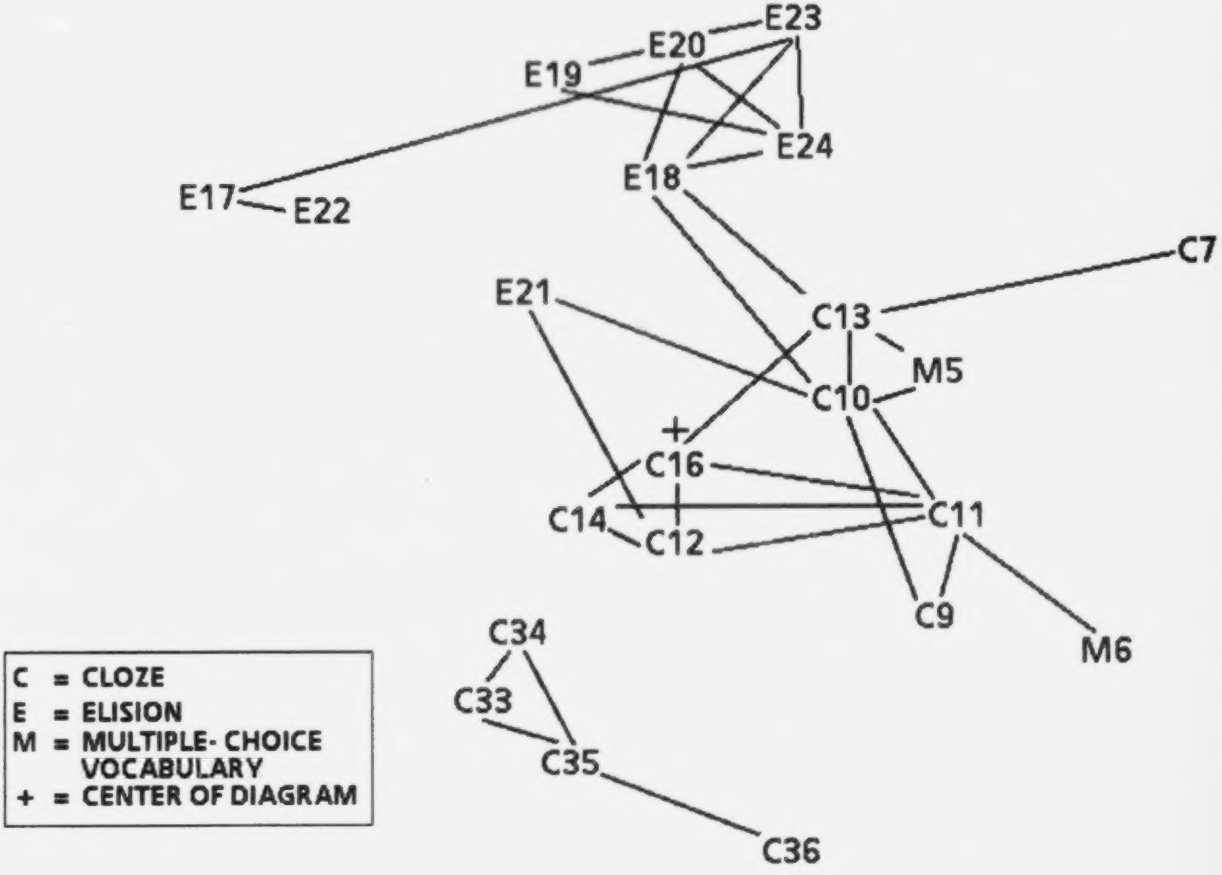

E23
E20
E19
E24
E18
E17
E22
C7
E21
C13
M5
C10
+
C16
C14
C12
C11
C9
M6
C34
C33
C35
C36
C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

Table 4b

THREE DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

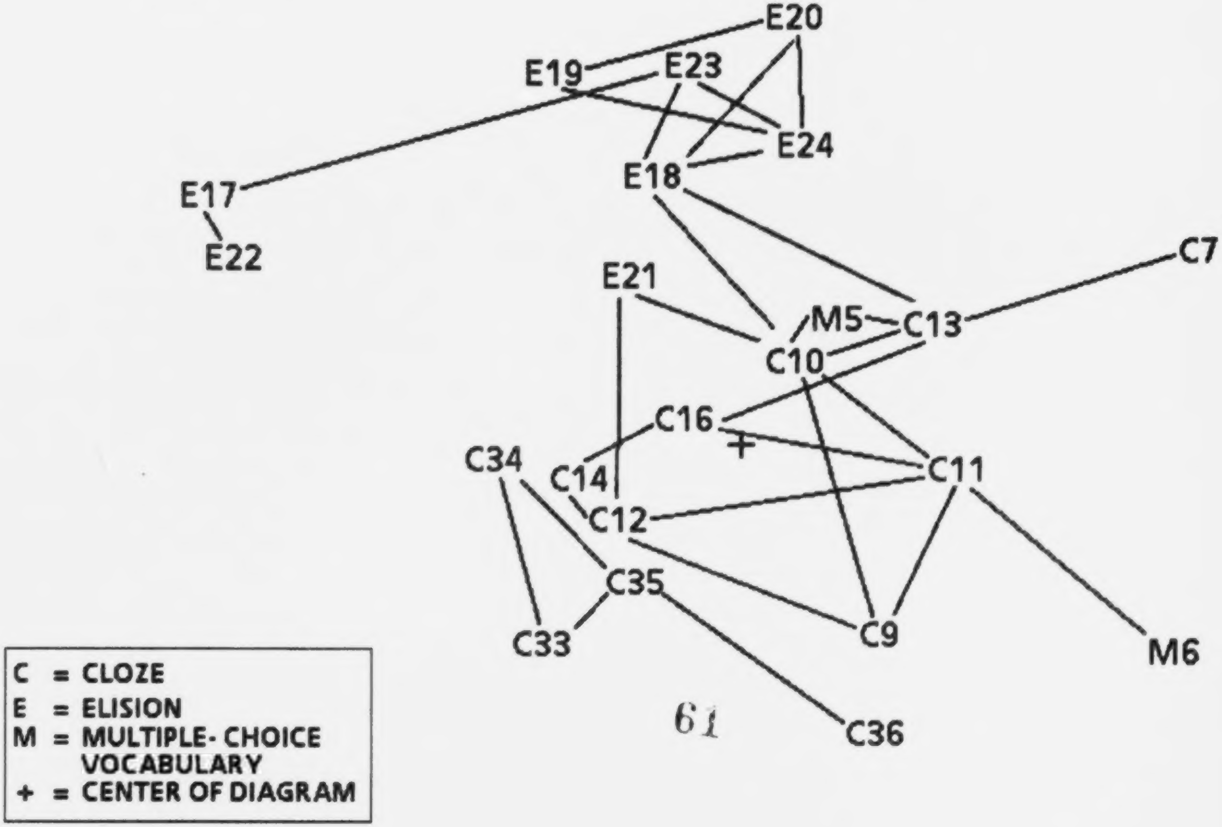

E20
E19
E23
E24
E18
E17
E22
E21
M5
C13
C7
C10
C16
+
C34
C14
C11
C12
C35
C33
C9
M6
C36
C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

Table 4c

FOUR DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

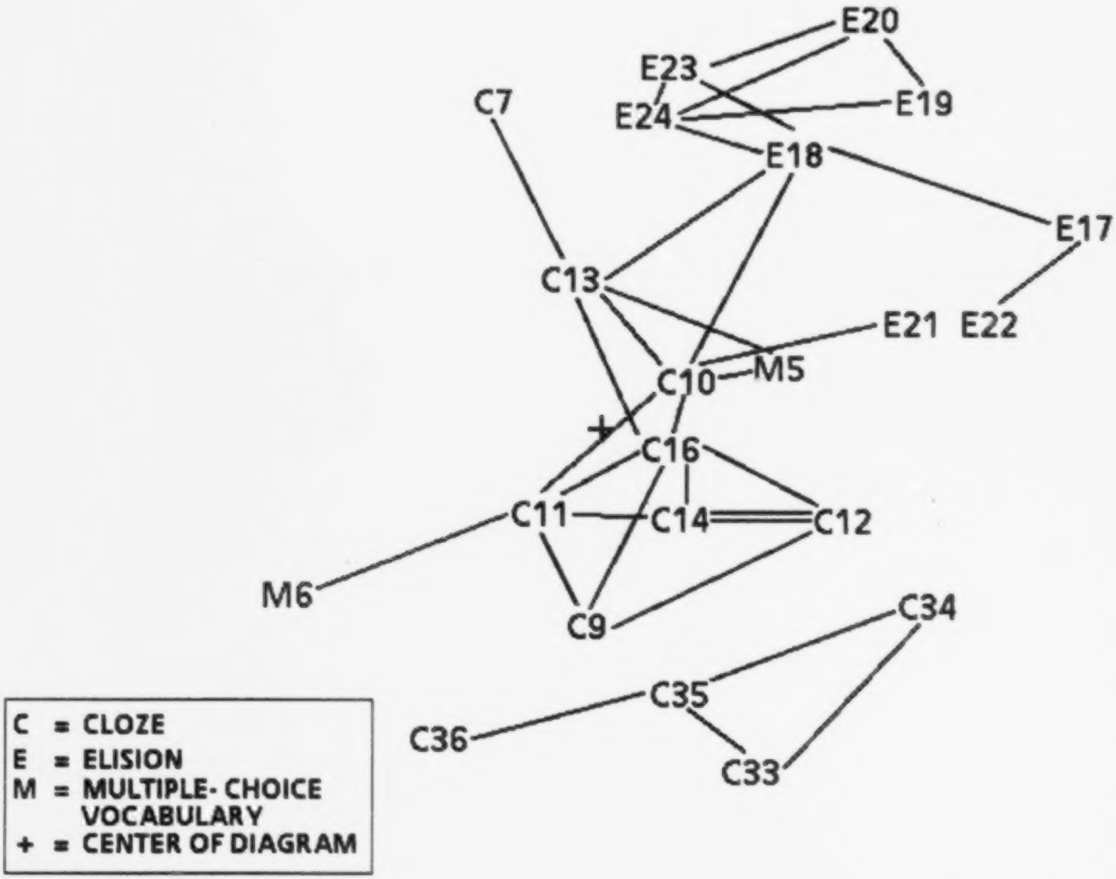
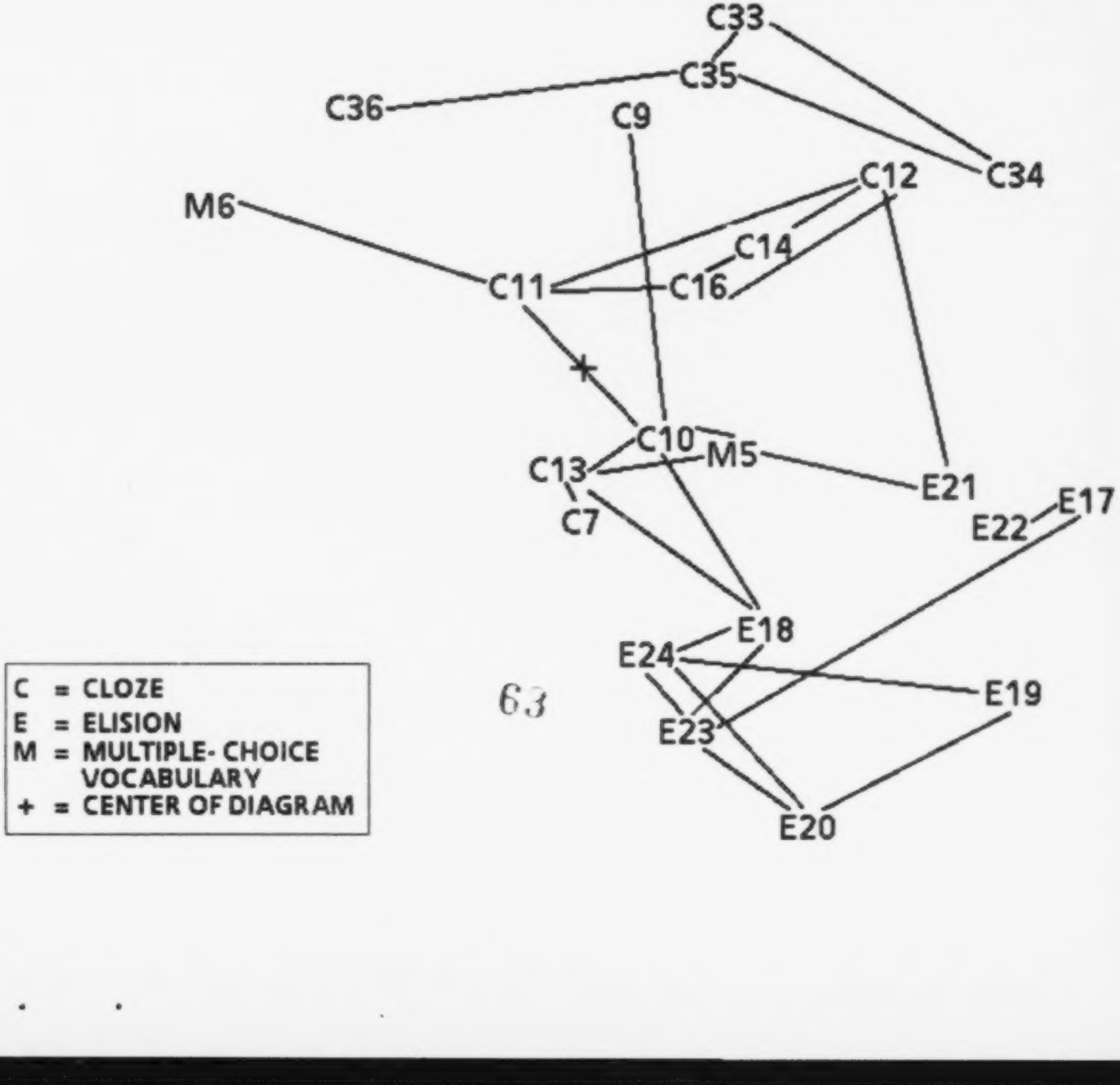

E20
E23
C7
E24
E19
E18
E17
C13
E21
E22
C10
M5
+
C16
C11
C14
C12
M6
C9
C34
C35
C36
C33
C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

Table 4d



FIVE DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

C33
C35
C36
C9
C34
C12
M6
C14
C11
C16
+
C10
M5
C13
E21
E17
C7
E22
E18
E24
E19
E23
E20
C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

AN EXPERIMENT WITH A NARRATIVE DISCOURSE TEST

Steven Ross
Osaka, Japan

With approximately forty-three percent of Japan's high school graduates going on to higher education, the study of foreign languages in Japan has reached a level of development comparable to the country's post-war economic growth. An important reason for the steady growth in foreign language education has been the college entrance examination. English as a foreign language is one of the essential parts of the national university standard examination, and is often the most important aspect of entrance examinations in the humanities departments of private universities and liberal arts colleges.

While the emphasis on English as a foreign language has grown, the basic methods of testing have not changed. Grammatical analysis and linear translation are still the preferred methods of evaluation, with little emphasis on listening comprehension and none at all given to speaking skills. These methods of testing have had a large impact on curriculum design in Japan, and the average college freshman is likely to have studied English as a foreign language for at least six years without ever having an introductory course in spoken English.

By the time the high school graduate has successfully entered a university, his or her first oral English course may be offered. A typical introductory course might be taught by a native speaker to a class of thirty to forty students. The disproportionate number of students to teacher in these college level classes precludes any serious attempt to test speaking proficiency directly. Often there are no provisions made to assess speaking proficiency at all at the university level unless an individual can pass qualifying examinations--again examinations that measure analytic ability and not oral skills.

Such examinations are commonly set by publishing houses or professional testing services. By far the most popular college level test, produced by the Society for Testing English Proficiency (STEP), does not directly measure speaking ability in the one-to-one interview format common in Britain and North America. Because of the disproportionate number of applicants to native interviewers, groups of students are typically confronted and asked to speak in turn on topics chosen by the interviewer, or on topics chosen at random from a pool of items. This approach results in a highly selective pre-oral examination screening and a method of speech elicitation that does not provide much more of a profile of an individual's sociolinguistic or discourse competence than would a conventional paper-and-pencil test of the individual's linguistic competence. Given this state of affairs, a practical method of oral language proficiency measurement which can reliably assess a foreign language speaker's ability to narrate under a reasonable time constraint can potentially serve to provide a more direct appraisal of the speaker's linguistic proficiency. Such a test could replace the paper-and-pencil screening/placement batteries and provide a more face valid pre-interview criterion.

Although indirect assessment of oral skills has been in use for some time (Lowe and Clifford 1980), the use of tape recorded question and answer formats in university-level oral skills assessment is now gaining popularity (Clark 1986). In Japan, where video hardware is readily available at virtually all institutions of higher learning, indirect assessment of oral skills has new potential in combination with voice recording equipment. For large-scale institutional testing, however, there are of course certain constraints which would apply to an assessment of a foreign language speaker's ability to narrate a short story utilizing the video as the medium of speech elicitation. Some of these are outlined below.

* Individual speech samples should be recorded on tape.
* Speech samples should be reliable.
* The speaking task should have multiple alternate forms.
* Language laboratory (L.L) assistants should be able to administer the test.
* Mass administration should be possible in a video-equipped L.L.
* Short random samples of the each recording should be scorable.
* Both holistic and atomistic scoring methods should be possible.
* Scores from the narrative discourse test should have high concurrent validity with face-to-face interviews.

On-Line Discourse

In the past decade only a few studies have explored phenomena associated with the production of narrative discourse (Chafe 1979, 1980; Ochs 1979). In a study involving the on-line narration of a short animated cartoon, Tomlin (1984) found that both native speakers and advanced non-native speakers were inclined to use a pragmatic mode of discourse which relied on left-dislocation and linear coordination of clauses, and in the case of non-native speakers of English, the non-use of grammatical morphology. Tomlin's study suggests that the use of a 'cold turkey' approach to narration, one in the unrehearsed extemporaneous mode, will cause a regression to a basic paratactic strategy by any speaker. In light of Tomlin's findings, such an approach to narration for the purposes of foreign language assessment would not be indicated since there would be an even more profound reduction of variance, especially if the sample of foreign language speakers to be tested is at the lower end of the ability scale.

In order to make oral narration a practical tool for foreign language speaking assessment, certain modifications to the extemporaneous production approach are necessary. Pre-viewing of the animated sequences in the learners' first language, for example, would allow each individual an equal starting point in terms of his or her anticipation of the outcome of the story line. The presentation of the complete story line, in the case of the pilot version of the project (an animated Japanese folk tale), allows each individual an equal opportunity to approach the speaking task in a culturally familiar framework.

One problem associated with the pre-viewing phase of the presentation of the narrative material arises, however, when individuals assume that they are to relate the subsequent scenes of the story from the perspective of each participating character. The resulting first-person narrative would be problematic in that the syntax manifest in its production is less elaborate than a third-person narration of the story line. This, then, necessitates a pre-viewing presentation of an example of a third-person narration of an animated cartoon.

The next phase, by far the most challenging aspect of creating a video narrative task, is in selecting and editing scenes of equal length and salience so as to maximize the potential for differences in speaking skill to emerge. This editing phase requires an optimal balance between the need to present the narrative task so that production would be in 'real time' and at the same time allow the speaker to anticipate the up-coming sequence of events in the story line and plan her discourse accordingly.

Proficiency Scales

The primary use of the narrative discourse task in this study was for screening and placement into a five-level course in spoken English for college freshmen. Consequently, the anticipated ranges of proficiency levels were limited to known parameters within the institution: from persons with virtually no experience speaking a foreign language to persons with considerable experience in acquisition-rich environments. No need was seen, however, to establish a native speaker norm for the holistic proficiency scales. Instead, a small sub-set of learners with extensive overseas contact with English were chosen as the anchor population to which the scales could be developed in a descending order. The result of this approach to scaling was a set of anchor scores in three components of basic speaking proficiency: pronunciation, accuracy and fluency. The descriptors for each of the subscales follow:

Pronunciation	"4"	Speaker can use reductions variably. Sentence level stress and accent are accurate.
	"3"	Speaker primarily uses English syllable structures. Stress and accent are variable.
	"2"	Speaker relies on Japanese syllable structure. English vowels are more prevalent than Japanese cardinal vowels.
	"1"	Speaker uses Japanese syllable structure and cardinal vowels.
	"0"	No response.
Accuracy	"5"	Speaker uses a relatively wide range of syntactic patterns and uses accurate verb morphology.
	"4"	Speaker can use a variety of syntactic patterns. Salient morphology is accurate.
	"3"	Speaker can provide a canonical sentence pattern for each scene. Verb morphology is unstable.

	"2"	Speaker must string canonical sentences across more than one scene. Little verb morphology appears.
	"1"	Speaker can only manage to provide isolated words.
	"0"	No response.
Fluency	"6"	Speaker can narrate the story at the same pace the scenes change, and can anticipate up-coming frames.
	"5"	Speaker can produce two or more T-units for some scenes. Multi-clause T-units appear.
	"4"	Speaker can manage a short T-unit for each scene.
	"3"	Speaker can complete T-units selectively. T-units may be started for each scene, but only longer scenes get completed T-units.
	"2"	Speaker can complete T-units by spreading them across two or more scenes.
	"1"	Speaker attempts utterances, but commonly abandons them when the scenes change.
	"0"	No response.

In addition to the above holistic ratings,[1] T-unit analyses were done on one hundred forty-eight of the recordings. Such a syntactic analysis of the recorded narrative was thought to be a viable method of determining the extent to which raters could differentiate between the accuracy and fluency criteria established in the holistic rating scales. An analysis which could provide empirically defined measures of accuracy and fluency would also give ample convergent and discriminant evidence for the holistic scales in that the atomistic T-unit analyses would serve to provide specific criteria with which comparisons of the holistic scales could be made. To this end T-unit analysis (see Gaies 1980, Larsen-Freeman 1978) was chosen because it is relatively transparent and does not require a great deal of rater training. However, the primary weakness in performing a T-unit analysis on samples of spoken narrative is the fact that the analysis of syntactic development begins at the clause level. Other approaches, such as that endorsed by Ferguson (1980), include sub-clause level criteria within tone groups as the primary units of analysis. A preliminary comparison of tone group analysis and T-unit analysis on the initial set of samples for rater norming, however, suggested that tone groups corresponded to T-units at the clause level.[2] Given that the focus of the syntactic analysis was to establish operationalizable criteria for fluency and accuracy, no precision was considered lost in beginning with T-units as the basic unit of syntactic measurement. The three primary categories of T-units of interest in the analysis were total spoken T-units, total error-free T-units and words in error-free T-units.

---

[1] All holistic ratings of pronunciation, fluency, and accuracy, as well as the atomistic T-Unit analyses, were done on recordings of the student narratives.

Reliability

Two hundred and forty-seven recordings of the first of two parallel forms narrative tests were collected for the initial analysis. Random samples of ten narratives were collated on single cassettes and distributed to four raters for the primary inter-rater reliability estimation. Once a reasonable coefficient of reliability was established on the holistic rating (W = .85), the T-unit reliability phase could be undertaken. In all, three sets of ten samples each were rated before the holistic rating of the entire corpus of recordings was started. Owing to the time required to complete careful syntactic analyses of the narratives, a smaller sub-set of one hundred and forty-eight comprised the corpus for the more detailed analysis. The time spent in the initial rater norming phase turned out to be well spent. The Kendall's concordance coefficient for the four holistic raters reached .85, and the objective scoring required in the T-unit analysis by three raters was greater than .90 for counts of total T-units and greater than .95 for the error-free T-units.

Factor Analysis of Sub-Scores

Two versions of the narrative discourse task were prepared. The first, based on a Japanese folk tale about "the cunning of foxes", was presented to the incoming freshman class at the beginning of the academic year. The second version of the task, "the magic carp," was presented to the same students at the end of the nine-month academic term. Both versions were presented in a 'state of the art' language laboratory equipped with large video screens at the front of the laboratory and individual tape recording facilities in each cubicle. Since 247 students were to be tested for each version, six rotations in and out of the language laboratory had to be managed in a four-hour period. In both administrations, totaling four hundred and ninety-four recordings, only one tape was lost to a mechanical malfunction.

---

[2] In the framework introduced by Ferguson (1980), sub-clause level analysis of tone groups would assign weighted scores to well-formed verbs or nouns with functional words attached within the tone group. The T-Unit approach used in the present study was limited to dichotomous categorizations of T-Units (clauses) as +/-error-free. The tone group method would thus assign sub-scores for accuracy for a sentence like "I go/to the bank/this morning." In the T-Unit approach, however, such an utterance would be considered a single spoken T-Unit.

Immediately prior to the viewing of the animated cartoons, all students were given a short battery of paper-and-pencil tests. Listening cloze, dictation and a multiple choice grammar test were selected from a larger battery of tests developed in previous years. These tests were included for the purpose of comparison with the scores derived from the speaking tasks.

Two separate correlation matrices were generated from data derived from the two administrations of the paper-and-pencil tests and the narrative discourse tasks. Specifically, input to each matrix included scores from the speaking task: ratings for pronunciation, accuracy and fluency; objective measures of total T-units, error-free T-units, words within error-free T-units, the ratio of words in error-free T-units to total error-free T-units (WEFTU/EFTU), the ratio of words in error-free T-units to total T-units (WEFTU/TU), the ratio of error-free T-units to total T-units (EFTU/TU), and the ratio of words in error-free T-units to the number of error-free T-units times the total of T-units in the narrative (WEFTU/EFTU × TU). Listening cloze, dictation and multiple choice grammar scores represented the paper-and-pencil method of testing. Each matrix was entered into a separate principal factor analysis and rotated to a three factor solution. The purpose of the separate analyses was to establish the similarity between the two separate speaking tasks after the nine-month period of instruction.

Table 1

Factor Structure of Two Narrative Discourse Tasks

Factors	Eigenvalues		% of variance		Accumulative % var.	
	pre	post	pre	post	pre	post
1	7.03	7.25	66.6	67.2	66.6	67.2
2	1.71	1.50	16.2	13.9	82.8	81.1
3	.76	1.04	7.2	9.6	90.1	90.8
4	.50	0.66	4.7	6.1	94.8	97.0

Table 2

Orthogonal Rotation

	Loadings **factor 1**		**factor 2**		**factor 3**	
	pre	**post**	**pre**	**post**	**pre**	**post**
pronunciation	.612	.496				
fluency	.855	.827				
accuracy	.672	.592		.577		
holistic total	.853	.744		.511		
dictation						.693
grammar					.446	.536
cloze					.445	.707
T-units	.887	.844				
error-free TU	.576	.492	.765	.855		
words in EFTU	.501	.417	.767	.855		
WEFTU/EFTU			.885	.941	.390	
WEFTU/TU		.413		.309	.665	.381
EFTU/TU			.939	.956		
WEFTU/EFTU x TU	.707	.745		.349		

Similar factor structures emerged on both versions of the narrative discourse test. Of main interest are the first two factors which account for about 82% of the variance on the two tests. Factor one consists of a single 'pure' measure of verbal fluency and complex measures that straddle the main fluency measure and the error-free measures. Factor two, whose most consistent measures are the ratio of words in error-free T-units to error-free T-units and the ratio of error-free T-units to total T-units, may be more a measure of morphosyntactic accuracy, given the fact that the measures containing the 'error-free' criterion load on this factor.

The holistic ratings partially support this dichotomy. The accuracy rating on the post-test loads with the objective accuracy measures as we might expect. One reason why the accuracy rating appears complex, with loadings on both factors on the post-test while only loading on the first factor on the pretest, may plausibly be traced to subtle changes in rating criteria by one or more of the raters. Once familiarized with the use of the objective measures of accuracy, i.e., those based on T-Unit analysis, the raters might have listened to the recordings counting error-free T-Units when they were rating the accuracy of the narratives on the post-test.

The fluency rating loads only on the first factor. The holistic total again loads with the first factor on the pretest and becomes more complex by the second administration of the narrative task, most likely because of the influence of the accuracy rating. The most striking feature of the holistic loadings, however, is the apparent inclination of the raters to recognize the gross verbosity produced by the speaker as a sign of a more general speaking proficiency. In contrast, there were a few students who remained relatively taciturn until they found a particular sequence in the narrative story in which they could produce a perfectly formed utterance, and it appears likely that the raters initially recognized their silence as an overt sign of lack of proficiency.

The paper-and-pencil tests did not load on the two speaking factors. Given the fact that the multiple choice grammar test could not be empirically parsed out from the two measures of listening comprehension, dictation and the listening cloze, it is possible that these six tests were a reflection of a method of measurement artifact.[3]

Concurrent Validity

One of the goals of developing the narrative discourse test is to find a task that will share substantial variance with a more direct interview task. To this end several interview formats were considered in the concurrent validation phase of the project. The use of 'quick and dirty' interviews (i.e., interviews with a modicum of face validity and little reliability) was a temptation given the large numbers of potential interviewees and the unavailability of a sufficient number of native speaker interviewers. However, a structurally-based face-to-face interview was chosen to avoid the 'clam-up' phenomenon common to first interviews with Japanese students. It was believed that a structural describe-a-picture type of interview would bypass this problem. Subsequently, the John Test (Language Innovations n.d.) was chosen as the interview method because it contained sub-sections on structure and a freer narrative description of a short story depicted on the student's test material. The John Test was administered to a small non-random sample of students placed into a pre-intermediate level EFL course. It is important to note that since the interviewed group had been streamed, the likelihood of a dispersion of interview scores was precluded. Nonetheless a product moment correlation of .77 (n = 44) was found between the holistic rating and the John Test interview. The correlation would no doubt have been larger had students from the entire range of courses been interviewed.

A secondary goal in devising the narrative discourse task is to filter out less viable methods of rating samples of speech and find a single rating scheme that will provide the greatest amount of information about speaking proficiency in the least amount of time. An essential step in finding consistent measures is checking the pre and post-test correlations. The table below suggests that the measures of accuracy were the most erratic, presumably owing to the effects of instruction during the nine-month period between test administrations. The fluency measures, on the other hand, show relatively more stability over time.

---

[3] The dictations used in this study were short (five sentences long) and had lead-ins to each of the sentences. The cloze passages, while much longer, involved listening to a reading of each passage and allowed ample time for the filling in of the gaps. It appears possible that, despite the aural presentation of the dictation and cloze tests, most of these Japanese students relied on a gap-filling strategy first instead of attending to the aural cues. If indeed this strategy was widely used, the distinction between the listening and grammar tests may have been diminished.

Table 3

Correlations Between Pre and Post Test Measures

pronunciation	.284	T-units	.564	cloze	.484
fluency	.490	W/E × T	.595	dictation	.315
accuracy	.445	WEFTU/EFTU	.244	grammar	.489
holistic total	.509	WEFTU	.442	EFTU	.334

(p <.01 for all correlations)

Conclusions

Given the impracticality of oral interviews in most foreign language testing contexts in Japan, screening for face-to-face interviews and placement tests can be facilitated with the use of narrative discourse tasks which can be recorded for analysis at a later date. In devising the tasks, care must be taken to provide appropriate background information to all of the examinees while creating speaking tasks that stimulate the speaker to narrate in 'real time'. The ultimate proof of the usefulness of the narrative discourse test, however, will be in its practicality; tape samples will by necessity be short and the rating criteria will need to be made more concise before this approach to testing oral proficiency can gain wider acceptance. The final phase of the experiment with the narrative discourse test will involve generalizability studies of progressively shortened samples of student speech so that the optimal intersection of reliability, validity and practicality can be estimated.

References

Chafe, W.L. 1979. The flow of thought and the flow of language. In T. Givon (Ed.), Discourse and syntax: syntax and semantics. 12, 159-182 New York: Academic Press.

Chafe, W.L. 1980. The pear stories: Cognitive, cultural and linguistic aspects of narrative production. Norwood, New Jersey: Ablex.

Ferguson, N. 1980. The Gordian Knot. Geneva, C.E.E.L.

Gaies, Stephen, J. 1980. T-unit analysis in second language research: applications, problems and limitations. TESOL Quarterly 14: 1.

Larsen-Freeman, Diane 1978. An ESL index of development. TESOL Quarterly 12: 4.

Ochs, E. 1979. Planned and unplanned discourse. In T. Givon (Ed.), Discourse and syntax: Syntax and semantics. 12, 51-80 New York: Academic Press.

Tomlin, R.S. 1984. The treatment of foreground-background information in the on-line descriptive discourse of second language learners. Studies in Second Language Acquisition. Vol 6 No. 2., 115-142.

The John Test (n.d.) Language Innovations Inc. 2112 Broadway Room 515. New York, N.Y.

Some Problems Concerning the Interpretation of Passage Correction Tests

Terence Odlin
Ohio State University

Passage Correction tests (PC) are measures which require individuals to identify and correct errors that have been inserted in a prose passage. There is a family of editing test formats that could fit the above definition, but a distinctive characteristic of the tests to be described herein is that no errors are overtly marked. When individuals have no warning about what in the passage is anomalous, they may and often do overlook errors. Such systematic oversights can be a valuable source of data for second language acquisition researchers and for teachers of writing.

Earlier studies of PCs by Davies (1975), Bowen (1978), Arthur (1980), and others have shown that PCs have some interesting similarities and differences with other language tests. The similarity of passage correction to the editing of one's own writing makes the PC attractive for programs teaching basic writing, and university researchers in England, New Zealand, and the United States have either used or have considered using PCs in their ESL programs. Nevertheless, passage correction tests have not attracted nearly as much attention as cloze tests or other measures, and so it is natural that PCs are much less understood. PCs deserve to be better understood, however, since they have considerable potential to produce interesting data on language transfer, monitoring, and the development of basic writing skills. Accordingly, this paper deals with results from two PCs that I have been studying over the last five years. The results suggest problems of interpretation that may arise with a wide variety of PCs.

The two tests to be described involve essentially the same task: students in ESL courses had to detect different types of errors and correct them. However, in other details the two measures differ considerably. One of the PCs, hereafter referred to as the Picture Test, was about two hundred words long and was accompanied by a picture. Test takers had to consult the picture in order to correct semantic inaccuracies as well as other types of errors (Odlin 1986). The other test, hereafter referred to as the Fossil Test, was about a thousand words long and was not accompanied by any picture. The text of the Fossil Test was a description of fossils that was written for a non-specialist audience having little training in science (cf. Odlin 1985). The Fossil Test was considerably more difficult than the Picture Test but, as the discussion will show, it has proven useful as a measure of performance in basic writing courses. Three issues relevant to one or the other (or both) of the tests will be discussed in this paper: 1) the use of native speakers of English as a baseline for gauging the performance of ESL students; 2) the comparative results obtained on the Fossil Test and on holistically graded essays; and 3) some results of item analyses.

Before these three issues are discussed, it will help to identify some general results from the two PCs. In a number of respects, the track records of both tests are good. The reliability coefficients (KR-20 and test-retest) have generally been between .80 and .90, a fact which is especially significant in view of the small number of items on both tests (26 on the Picture Test and 24 on the Fossil Test). Another characteristic of both measures is that the results correspond to results from other types of placement tests. In the case of the Picture Test, the results were compared with those on the Comprehensive English Language Test (CELT) and there were correlations between .83 and .87 ($p<.01$ in all cases) when different scoring systems were used. The Fossil Test results, which will be discussed in detail below, show a stratification similar to that obtained on a placement exam involving essays scored holistically. Moreover, both tests show that native speaker control groups outperformed ESL students.

In contrast to these similarities between the two PCs, there are test-specific characteristics that are noteworthy. The Picture Test showed significant correlations between sub-types of items that were semantically or grammatically related in different ways, and results suggested that individuals sensitive to certain types of semantic anomalies were likely to score high on the CELT. The Fossil Test indicated that students who received special training in error detection showed a significant score improvement while individuals who received no training showed no improvement. Such results suggest the potential of the PC format not only for further research in second language acquisition but for second language teaching as well. Using PCs well presupposes understanding them well, however, and it is thus desirable to consider now some of the factors related to interpretation of such tests.

The Native Speaker Baseline

The results from the Picture Test indicate a clear difference in native and non-native abilities. Table 1 shows that the control group of 20 native speakers had little difficulty with the test whereas the ESL students generally had great difficulty, no matter what their ability level was. While the native speaker scores are not perfect, the mean is nearly twice as high as that of the most advanced of the ESL groups. This result, along with the fact that there was no consistent pattern of items missed by the native speakers, makes it reasonable to attribute the less-than-perfect scores of the control group to oversights on their part.[1]

Table 1

Items Correct on Picture Test

	IEP 0-1	IEP 2-3	IEP 4-5	IEP 0-5	CONTROL
Mean	1.5	6.3	13.8	7.2	23.1
S.D.	2.4	3.7	2.8	5.8	2.0
N	8	9	8	25	20

---

[1] The less than perfect scores cannot be attributed to defective test items. Such items were excluded from the scoring (for details, cf. Odlin 1986).

In Table 1 IEP stands for Intensive English Program. The maximum possible score was 26.

The Fossil Test results also showed differences between native and non-native speakers, but there are some interesting wrinkles in these results. First, two different native-speaker groups that have taken the test have produced baseline scores that are significantly different. One group consisted of 30 individuals from Ohio State University who were enrolled in an upper-division class on traditional English grammar; the second group consisted of 35 freshmen at the University of Texas at Austin who were enrolled in two different sections of an English composition course. The composition course was not, technically, a remedial writing course. However, many of the students in the course had been only conditionally admitted to the University, and their instructors confirmed what I had heard from other instructors teaching the same course: namely, that the provisionally admitted students were considerably weaker in English than were freshmen who had been admitted unconditionally. In contrast, the English grammar class consisted largely of juniors and seniors with some graduate students, and in some cases these individuals were English teachers or teacher-trainees. Not too surprisingly, then, the students in the grammar course scored higher than those in the composition course, with the scores being significantly different on a t-test, as shown in Table 2.

Table 2

Native-speaker Scores on Fossil Test

	AGC	FCC
Mean	17.0	13.2
S.D.	3.0	2.9
N	30	35

Maximum possible score = 24; $t = 5.63$, $p < .01$ (two-tailed)

FCC = Freshmen composition class; AGC = Advanced grammar class

These results led to a number of questions:

1) Could the differences in native-speaker performance be due to some difference in linguistic proficiency? The answer most probably is "No." Only individuals who were native speakers of English were included in the scoring, and there are no indications that regional dialects played any role in the results.

2) Are differences in metalinguistic awareness involved? Here, the answer most probably is "Yes." The differences between the two native speaker groups already described strongly suggest that the students in the grammar course had a wider-ranging awareness of linguistic forms and functions.

3) Do the less-than-perfect scores result from inappropriate items? Here, the answer most probably is "No." When individuals failed to notice an error,

the reason could generally be attributed to the location of the errors--either the location within sentences or within paragraphs. More details on the locations of errors will appear below.

4) Did individuals have enough time to complete the test? Here, the answer is less certain. For the sake of uniformity in testing conditions, all groups taking the Fossil Test have been limited to 30 minutes. It is likely that extra time would have improved some scores. However, 30 minutes did prove to be sufficient for some individuals to obtain near-perfect scores. In addition, many native speakers succeeded in identifying 24 real or imagined errors in the text, and so extra time would probably not result in significant changes in the scores of such individuals.[2]

Comparative Results of PCs and Holistic Evaluation

The results to be discussed in this section come from a comparison of student performance on the Fossil Test with student performance on essays scored holistically by staff in the ESL writing program at Ohio State University. While the PC results generally support the validity of holistic evaluation, they indicate some potential problems with this scoring procedure.

The writing program consists of three courses, 106, 107, and 108, entry into which is initially determined by an essay examination that is scored holistically.[3] All exams are read by two program staff members, who invariably have had considerable training and experience with holistic grading, and in cases where their placement recommendations do not coincide, a third reader is consulted. For foreign students, admission to Ohio State requires a minimum TOEFL score of 500, and most of the students taking the essay exam have TOEFL scores between 500 and 600. Students placed into a 106 course normally take 107 and 108 later the same year, and their progress in each course is judged as satisfactory or unsatisfactory largely on the basis of their performance on essay exams given at the end of each course. These final exams are also scored holistically, again by two readers with a third reader being consulted in cases of disagreements.

---

[2] One frequent and interesting result of PC tests is the fact that some individuals go on "witch hunts" of imaginary errors in sentences that are acceptable to most native speakers as well as to linguists. On the Fossil Test, for example, several native speakers of English made changes in the sentence I felt my horse to be on solid ground, a sentence which shows Subject-to-Object raising. Noonan (1985) has noted that raising is by no means a universal phenomenon among languages, and Kellerman (1983) has observed that Dutch EFL students will deem sentences with raised constituents to be less grammatical than sentences without such contituents. Native speakers of English may also have such intuitions about raising when certain types of predicates are involved.

[3] Actually, the Ohio State ESL writing program also divides courses into graduate and undergraduate sections, but those divisions are not relevant to the analysis at hand.

The Fossil Test results to be discussed come from the performance of six classes, two at each of the three levels described above. The results, which are presented in Table 3 in terms of the three course levels, show a clear trend: students in 106 found the PC much more difficult than students in 107 and 108 did. Confirmation of this trend is seen in the F statistic from a one-way analysis of variance ($F = 14.46$, $p < .01$). Nevertheless, there is an obvious discrepancy in the trend: While the difference of means between the 106 and 107 groups is large (and significant at the .05 level on a Tukey multiple comparison test), the difference of means between the 107 and 108 groups is small (and statistically non-significant).

Table 3

Items Correct on PC

	106	107	108
Mean	5.0	9.3	10.1
S.D.	2.3	3.4	4.0
N	21	24	21

One-way ANOVA results: $F = 14.46$, $p < .01$; Maximum possible score = 24

The negligible difference between the 107 and 108 groups might suggest a ceiling effect on the PC. However, results from other groups taking this test suggest that it is indeed possible for non-native speakers to score higher than 10 on the Fossil Test within the 30-minute limit. Consequently, other explanations for the flattening of the linear trend must be sought.

The most likely explanation is that there are only small (if indeed any) differences between the 107 and 108 groups, at least with respect to whatever skills PCs and essay tests both measure. If this explanation is correct, then the PC provides a more accurate indicator of those skills than do the essay tests. The evidence to be presented does support this interpretation, but it also suggests that there are some discrepancies between the two sets of holistic evaluations (for placement tests and for final exams). These discrepancies suggest why the PC and the essay tests give two different pictures of writing ability among 107 and 108 students.

Relevant to these discrepancies are some differences in the student populations in the 106, 107, and 108 levels. All students in the 106 class had been placed there as a result of the holistic score assigned to their placement exam. In contrast, 16 of the 24 107 students had done well enough on the placement exam to be exempted from 106, while 8 other students had previously taken 106 and had done sufficiently well on the final exam to move into 107. These 8 students will be termed "107 move-throughs." Of the 21 students in 108, only four had been placed there as a result of the holistic score on their placement exam, while the others had already taken 107, had passed the final exam, and had moved into 108. These students will be termed "108 move-throughs." (See Brown, 1981, for a similar analysis.)

Table 4

Perforance by 107 Students on PC
as a Function of Entry into 107

	106M	107P
Top half	2	10
Bottom half	6	6

106M = Previous enrollment in 106
107P = No previous enrollment in 106
Chi-square = 4.69, p.<.05

The proportions of 107 and 108 move-throughs within their respective classes are clearly different, and a natural inference is that there is at best only a small difference between students enrolled in 107 as a result of their placement exam scores and students who have moved into 108 on the strength of their performance on the 107 final exam. In other words, it may be the case that there is no significant difference between certain writing skills of 107 and 108 students. Corroboration for this inference comes from a comparison of the 107 students who were move-throughs from 106 and those students enrolled in 107 on the basis of their placement test scores. Table 4 classifies all 24 107 students into the half that scored above the median 107 score on the PC (which was 9.5) and the half that scored below the median. It is clear that a higher proportion of 107 move-throughs are in the bottom half and that a higher proportion of students originally placed into 107 are in the top half. The proportional differences are significant on a chi-square test (4.69, p<.05), and the evidence thus suggests that the PC results have to be interpreted in light of how students came to be in 107 and 108.

Such an interpretation is tantamount to saying that the two holistic scoring procedures, for the placement test and for the final exam, do not really give the same results even though the procedures superficially seem the same. A student essay written for a final exam appears more likely to elicit a more favorable evaluation than the same essay would if it were written on a placement test. This discrepancy in evaluations probably has two sources. First, the final exam measures performances of individuals considered to have somewhat comparable writing abilities, whereas the placement exam measures performances of individuals considered to have a wider range of abilities (from 106 to 108). A natural outcome, then, is that the 107 performances on the final will seem more homogeneous. In fact, the grading patterns in the ESL program support this interpretation: very few individuals have stood out either positively or negatively such that few individuals have received either A's or failing grades. Thus most individuals in 107 would be likely to pass, albeit without flying colors.

and to enroll in 108.[4] The second explanation is that the raters of final exams often have less training and experience than do the raters of the placement exams. It is thus possible that evaluations of the final exam are less dependable because of the relative inexperience of some of the graders. There is in fact anecdotal evidence supporting this explanation although space does not permit discussion of that evidence.

Relationships Among Individual PC Items

The final problem to be discussed involves item analysis. The specific data to be considered come from one administration of the Fossil Test at the University of Texas. The internal consistency of the test as measured by KR-20 is moderately high: .80. At the same time, however, the inter-item correlations are generally low and frequently non-significant, even in the case of correlations between items involving the same structural error. For example, only one correlation is significant between three items involving errors of third-person-singular in the present tense (e.g., Such rock often show many fossils...), the three correlations being .07, .09, and .61. It seems likely that a number of factors must be considered in determining why individuals who can detect one error involving the third-person-singular do not detect the others. Differences in the sentence structures within which the errors appear are one possible explanation. However, another factor that probably plays a larger role is the fact that some individuals concentrated their detection efforts on earlier parts of the text while others focused on the last two pages, where two of the three errors were. (The focus of such individuals is evident from the fact that on several individuals' papers many errors were often marked on one page with few errors being marked on other pages.) Aside from these differences of focus, there are quite possibly other types of reading strategies which could strongly influence what one detects, especially in a long text such as the Fossil Test. This is clearly one area where more research on PCs would be useful.

Yet while the inter-item correlations are low, the KR-20 suggests there is considerable internal consistency in the results. The KR-20 estimate is corroborated by the results from 23 multiple regressions in which each test item was the dependent variable and the remaining items the independent variables.[5] The R-Square coefficients obtained from these regressions ranged from a minimum of .50 to .97 with the mean R-Square being .76. Thus the regression equations were normally able to predict about 76 percent of the variance on an item. In other words, performance on any given PC item was generally predicted rather well by collective results from every other item.

---

[4] There have been recent changes in grading practices so that some of the observations made here about the writing program are now purely historical.

[5] One item with a variance of zero was not included in the regression analyses.

References

Arthur, B. 1980. Gauging the boundaries of second language competence: A study of learner judgments. Language Learning 30, 177-195.

Bowen, J. D. 1978. The identification of irrelevant lexical distractions: An editing task. TESL Reporter 12 (1), 1-3, 14-16.

Brown, J.D. 1981. Newly placed students versus continuing students: Comparing proficiency. In J.C. Fisher, M. A. Clarke, and J. Schachter (Eds.), On TESOL '80, building bridges: research and practice in teaching English as a second language. Washington, D.C.: TESOL.

Davies, A. 1975. Two tests of speeded reading. In R. Jones & B. Spolsky (Eds.), Testing language proficiency, (pp. 119-127). Arlington, VA: Center for Applied Linguistics.

Kellerman, E. 1983. Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), Language transfer in language learning, (pp. 112-134). Rowley, MA: Newbury House.

Noonan, M. 1985. Complementation. In T. Shopen (Ed.), Language typology and syntactic description, (pp. 42-140). Cambridge: Cambridge University Press.

Odlin, T. 1985. Passage correction as a measure of writing skills. Paper presented at the the 19th Annual TESOL Convention, New York.

Odlin, T. 1986. Another look at passage correction tests. TESOL Quarterly 20, 123-130.

PERFORMANCE PROFILES FOR ACADEMIC WRITING

Liz Hamp-Lyons
University of Michigan

One of the many innovative features of the British Council's English Language Testing Service (ELTS), introduced in 1980, was the inclusion of a direct test of writing. The context of the test is the testing of the English proficiency of overseas non-native English speaking, mainly postgraduate, students who are applying for scholarships to British universities and other tertiary education institutions, and who are normally applying for scholarships from the British Council or one of the agencies whose funds are administered by the British Council.

Although direct testing of writing was very common until the 1930's or 40's, and was indeed the only test method in the 1800's, in the structuralist-psychometric era 'essay tests' had fallen into disrepute and disfavor as unreliable. The emphasis on language as communication in the early 1970's and the humanistic trends of the late 1970's, however, were reflected in an emphasis on test validity, and led to a search for tests which would combine the essential qualities of reliability with validity. In addition, developments in ESP emphasized face validity among the other validities, and led to a particular interest in performance testing. The British Council's decision to include a direct writing test in the ELTS battery was, then, a logical part of a general pattern in language teaching and testing.

The ELTS writing test is the second part of the Modular section of the ELTS (hence the abbreviation 'M2'); there are six Modules, and the candidate takes whichever is most closely akin to her/his own field of specialization. I do not propose in this chapter to discuss the specific versus general issues at all, focusing instead on one part of M2 which is common across Modules, that is, the scoring procedure.

M2 consists of two compulsory questions, each based on an input text which the candidate has previously read in another part of the test. The test lasts 40 minutes, and the first question (recommended time 25 minutes) has come to be described as "divergent," in that it requires the candidate to consider the information in the input text in relation to the question, but also expects the candidate to bring a personal response to the question, for example by relating it to her/his own country or own special subject interest. The second question (recommended time 15 minutes) has come to be known as "convergent," in that it requires the candidate to stay very close to the input text, extracting and organizing the information to fit the needs of the question. (Questions are confidential and may not be divulged.) I will focus here on the development of the scoring procedures for the first question.

When M2 was first introduced, scoring was done with the aid of a short paragraph explaining the need to value communicative quality more than structural and surface features, but the main guide for the rater was a set of performance descriptions each associated with a performance level from 1-9, coupled with an example of performance at each level (i.e., a benchmark paper).

These nine levels and associated descriptions have become known as 'band descriptors', and I will refer to them throughout as the 'Original Assessment Scale', and to this scoring method as the 'Original Method.' The Original Assessment Scale is given in Appendix A.

It became clear that raters, most of whom were British Council ELT officers working in centers outside Britain and often in isolation from other M2 raters, needed firmer guidance in rating M2. An M2 Assessment Guide was written, piloted, revised and put into operation in 1985. This Assessment Guide took the criteria which had been implicit in the original general explanation of what should be valued in M2 answers and made these explicit. Each criterion was extensively characterized and some key problems raised by raters (e.g., "How long must an answer be before it can be looked at as 'communication'?" or "What constitutes plagiarism and how is it to be rated?") were tackled. The Guide took a self-study standardization approach, and included a set of 'criterion' papers for trial scoring, with discussion of how the standardization team handled them, and a further set for refresher scoring. The criteria for assessment of the first question presented in the 1985 Assessment Guide are:

Communicative quality
Organization
Argumentation
Linguistic accuracy and appropriacy

The rater is required to skim-read the essay three times: the first time, the rater focuses on communicative quality, i.e., a holistic reading, and makes a broad judgment which encompasses a three-band range (e.g., 2-5, 7-9, etc.). The second time the rater focuses on organization and argumentation and narrows the original judgment to a two-band range (2-3, 7-8, etc.). Finally, the rater reads again focusing on linguistic accuracy and appropriacy and decides on a single band from the two-band range, which is the final score for this question. This procedure is know as the 'Global Method.'[1]

But the Guide did not revise the original bands, and this presents several difficulties for the rater. First, the labels 'Competent Writer,' 'Marginal Writer,' etc., are difficult to interpret: in a trial with 20 raters I found that only 14 were able to correctly match the labels with the descriptions and number the sets in the correct order 1-9. My experience in training raters also suggests that the labels tend to discourage raters from looking closely at the full performance descriptions. But a more serious difficulty, and this is a recurring problem in rating essay tests, is that a single scale implies a unidimensional view of writing proficiency, and necessitates the treatment of each essay as existing at a single level. This would not be problematic if all writers did in fact write at a single performance level (thus manifest a 'flat' profile), but reports from raters indicated that sometimes a rater had problems rating an essay because she or he could not see one uniform level in the essay. Detailed observation of how raters actually rate revealed that the raters' instincts were accurate: the essays they had problems with were assessable using the criteria which had been developed, but only by looking at each criterion separately. The writers of such essays showed greater proficiency on some criteria than others. In ELTS terminology this multidimensional proficiency is referred to as a 'marked' profile. While the term was introduced to describe variations in proficiency across skills, it applies equally across the dimensions of a single skill such as writing.

---

[1] A copy of the relevant section of the M2 Assessment Guide may be obtained by writing to the Consultant, ELTS; ELLD, the British Council; 10, Spring Gardens, London, SW1A 2BN, England.

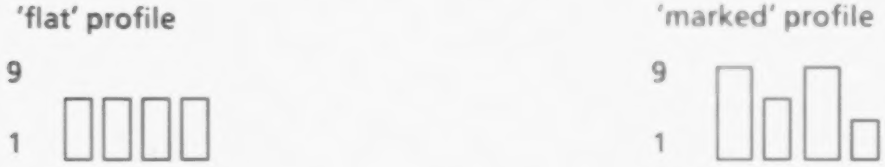


Figure 1: 'Flat' versus 'marked' profiles.

The Assessment Guide, therefore, incorporated two scoring methods: the unidimensional assessment (Global Method) is applied first, to all essays, while the 'Profile Method' (referred to later in this paper as 'Profile Method 2') was developed for use with, and only with, problem essays. At the heart of this method is the Profile Grid which schematises each criterion separately and provides a scale with the numbers of the bands for each of the criteria (Figure 2). Raters are asked to circle a three-band range on each criterion on the Profile Grid, and can then either choose the mode as their final score, or total the mid-bands on the criterion and divide by 5.[2]

As stated above, the intention was that the Profile Method would only be used with problem essays, after an initial application of the unidimensional assessment (Global Method). It became clear, however, that some raters began to apply the Profile Method to every paper. This meant that instead of looking first at communicative quality and then moving through organization and argumentation to linguistic appropriacy and accuracy (i.e., macro to micro features) in the Global Method as had been intended, these raters began with linguistic features and moved in the opposite direction. This resulted in more emphasis being placed on linguistic features than had been intended by the test design. We may speculate as to the reasons for this preferred use of the Profile Method: it may be because there is no such thing (or, at least, that some raters perceive no such thing) as a 'flat' profile, that is a writer whose proficiency is the same on every aspect of the writing process; or it may be that the Profile Method artificially creates multiple samples, permitting an objectivisation of what is for some raters an uncomfortably subjective process.

---

[2] A copy of the relevant section of the M2 Assessment Guide may be obtained by writing to the Consultant, ELTS; ELLD, the British Council; 10 Spring Gardens London, SW1A, England.

Practical use of the Assessment Guide, then, indicated that it had improved matters considerably, but that there was room for further development. In particular, there were three key reasons for revisions of the original assessment scale. First, the criteria were not fully or consistently articulated in the original assessment scale: the test designers had themselves been searching for a sense of what the criteria were or should be. This could only be known as a consequence of the operationalization of the test. Revision would permit the scale to be brought into line with the Assessment Guide as a whole with a clear and consistent treatment of the same criteria. Detailed observations of raters during the development and piloting of the Guide had shown that raters found their task much easier when the performance descriptions were presented as direct linguistic parallels, with the same criteria in the same sequence in each. Second, revision would permit the careful and clear differentiation of the nine levels of performance on each of the criteria. Raters had reported that they found difficulty differentiating between bands 6 and 5 in particular on the original assessment scale. Clearly, the consistent format also helped in this area. Perhaps most importantly for the long term, revision would allow the integration of the profiling principle which is at the heart of the philosophy of the ELTS, by taking account of marked as well as flat profiles; a first attempt had been made in this direction in the Profile Method, but as noted above this system had weaknesses.

Since the criteria had stood the test of practical use well, the first stage of the revision, the construction of a new set of global performance descriptions to match these criteria, was not difficult, although it required more than one piloting to ensure that raters could satisfactorily distinguish between the levels all the way along the scale. One significant change implemented in the Revised Assessment Scale was the separation of linguistic accuracy and linguistic appropriacy into two criteria. This was done for two reasons: first, some raters had reported occasions on problem papers when they felt there was a difference in performance for the same student on features of accuracy compared to features of appropriacy; second, it was generally agreed that the linguistic qualities of essays should be given more emphasis than they were in the Original Assessment Scale. This separation had already been implemented in the Profile Grid, but was now made more explicit. When the Revised Assessment Scale was ready (Appendix C), I found in a trial with 20 raters that they were all able to correctly sequence the nine levels (bands) without access to labels. (Note that in the Revised Assessment Scale labels are not used.)

The second stage of the revision was to develop a Profile Scale to enable each criterion to be examined independently. This would enable the handling of problem papers in the same way as had been done with the Profile Grid, although at a finer level of detail. This simply involved separating out the criteria of the Revised Assessment Scale and presenting them in individual columns. These are referred to as the new Profile Scale and Profile Method 2 (PM2). (See Appendix D.)

The rater is asked to choose a single band to describe performance on each criterion, not the three-band range previously used. It was believed that when criteria are sufficiently precise there is no reason to work from imprecise ratings, and that the imprecision had added to the difficulties of score aggregating in the first Profile Method. In other ways, however, the new Profile Scale (Profile Method 2) presents the same problems of score aggregating as the first Profile Method did. There is no mathematical formula for score aggregating;

combining scores on organization and linguistic accuracy and calling the answer writing proficiency is much like adding two apples and three pears and calling the result a lemon. Nevertheless, it has to be done, since clearly those responsible for absolute acceptance/rejection decisions for university places or for scholarships must have a single number to use. Whatever ethics or aesthetics may desire, this is the practical reality. It must be the test developers' responsibility to advise the score consumers of their best estimate of the candidate's writing proficiency, treating as unidimensional that which experience has shown is not unidimensional. The way in which the separate scores are aggregated must reflect the belief of the test developers about what is important in writing performance for the specific context, and in what proportions compared to other dimensions entering the same equation. There is no single 'right answer.' The answer which was arrived at for the particular ELTS M2 context was to weight communicative quality twice and all the other criteria once. However, no one involved believes this is an insignificant decision, and it is one which may be revised in the future as a result of the study of the test in operation, which is always continuing.

This chapter has so far focused on the validity of the assessment of the first question of M2: let us now consider reliability. In a small study comparing the various scoring procedures developed so far for M2, 12 inexperienced raters worked in four teams of three, each team using one of the scoring procedures. The raters were chosen mainly for their availability and willingness, but also as being suitable candidates for positions in British Council DTO's (Direct Teaching Operations, i.e., British Council centres where English is taught), and therefore potential raters of M2 in the field. The four scoring procedures used were:

1. original assessment scale using the original bands (OM;Appendix A);
2. original assessment scale, combined with Profile Method 1 (PM1; Appendix B);
3. revised assessment scale and the global method (RM; Appendix C);
4. new Profile Scale and Profile Method 2 (PM2; Appendix D).

All the raters were given the same general introduction and a short training session, rating two answers by their assigned scoring procedure. Each team then rated the same ten answers, first giving an individual rating and then agreeing on a final rating. Investigation of the scores assigned by the raters as individuals as compared to the scores assigned by raters as teams showed that the original method (OM) resulted in the largest number of rater disagreements (defined as each rater having a different score, i.e. at least a three-band spread for the three scores): raters disagreed on five out of ten answers; one answer received ratings of 7, 5 and 3. The original assessment scale combined with Profile Method 1 (PM1) resulted in two cases of rater disagreement, the revised assessment scale (RM) resulted in one case of rater disagreement, and the revised Profile Method (PM2) resulted in no cases of rater disagreement. The average single rater reliabilities for the four methods were:

OM	.563
PM1	.864
RM	.883
PM2	.942

Using the Spearman-Brown prophecy formula the reliabilities with three raters are estimated thus:

OM	.790
PM1	.950
RM	.960
PM2	.997

(It can be seen that with a more reliable method there is proportionately less additional reliability for more raters.)

On this preliminary check, then, the development of the new rating scale seems to have been of marked benefit to reliability. The use of a performance profile approach in the form of both the profile grid (PM1) and the profile scale (PM2) also contributes something to reliability: in the case of the addition of the first profile method to the original assessment, the result is a major increase in reliability; in the case of the addition of the second profile method to the revised assessment scale, the increase in reliability is only slight, and the single rater reliability for RM is, on this sample, more than adequately reliable.

In the British Council context, as explained above, the practical reality is that M2 is scored by a single rater, often working in considerable isolation from other raters. What must interest us in this context is a high single-rater reliability rather than any theoretically but not operationally achievable multiple-rater reliability. For this purpose any of the methods except the original one is acceptable.

The correlations between the four methods were generally quite high. Listed below are the correlations between the aggregate scores for sets of logical 'pairs:'

OM/PM1	.908
RM/PM2	.920
OM/RM	.845
PM1/PM2	.929

The other two correlations are .827 for OM with PM2, and .864 for PM1 with RM.

It can be seen that the highest correlation is for PM1 with PM2. These two methods are very similar in allowing the rater to treat each essay as a multiple sample: conceptually they share a view of writing as (at least potentially) multidimensional. We may hypothesize that the profile grid, although it was without any descriptions for the different criteria at each level, achieved what had been intended simply by allowing the rater the 'space' in which to respond. PM2 takes this much further than PM1, but it may be more an administrative convenience than anything else, since the descriptors are already present in the global version of the revised assessment scale: all the profile version of the scale does is break them up conveniently.

The high correlation for RM/PM2 is an important confirmation that these two related methods are yielding comparable scores, which is essential when two methods are used as alternate possibilities with the same set of candidates, as these two methods are. The correlation for OM/PM1, while not quite as high, is similarly at a reassuring level. We do not, of course, yet have data to show whether similar correlations will be achieved in the field, with real 'problem' answers.

It is worth noticing that the choice of scoring procedure appeared to have a slight but noticeable influence on the resulting score level: Table 1 shows that OM tends to be more generous than the other three methods, and that PM2 tends to be more stringent. It would appear that as the scoring method has been refined and become more rigorous, it has also become more stringent. When averaged, these differences are quite small, but on a single-scorer test any differences may be very dramatic for any one individual. It is, therefore, heartening that the trend in the development of the methods has been towards increasing reliability. Table 1 shows the aggregate score for each answer for each scoring method:

Table 1

Aggregate Scores: Essay x Scoring Method

	Method			
	OM	PM1	RM	PM2
Essay				
1.	6*	7	7	7
2.	7	7	7	6*
3.	5	6*	5	5
4.	6	5	6	5
5.	8	8	7	7
6.	4*	3	2	2
7.	4	4	5	5
8.	4	3	4	3
9.	6	6	5	5
10.	5	5	4*	5

The single asterisk (*) indicates where there is an aggregate score which is noticeably different from the others: however, even in these cases the 'wild' score is only different by a single band (e.g., No. 1:7*:6:6:6:). The widest range of aggregate scores on the different scoring methods is three bands (e.g., No. 6:4*:3:2:). Nevertheless, this is a significant difference if it is, for example,the difference between a band 5 and a band 7: band 5 is unlikely to be considered acceptable for scholarship purposes without remedial English, while band 7 is almost certain to be acceptable as it stands. It can be noted in passing that anecdotal reports from score consumers, such as tutors on EAP courses, suggest that the M2 score has often been found to be over-generous. The increasing stringency noted above may provide a more accurate reflection of candidates' writing proficiency.

It must be remembered that the raters used for this investigation were not well (or uniformly) trained or experienced. However, the brief training they did receive was with this researcher, i.e., an experienced trainer/rater who thoroughly understood each of the methods and their rationale. We cannot know for sure what use is made of the Assessment Guide by raters in the field. Only after the new methods have been in use for some time and sufficient data have been collected and analyzed will we know whether similar patterns emerge.

Finally, it should be noted that the real advantage of the profile methods lies in their diagnostic function, especially in the case of PM2: if we can achieve not only more accurate information, but more information, we open tremendous potential for the use of test results in other contexts, and the test instrument makes gains in practicality. A testing system such as the ELTS is predicated on the belief that by administering tests of different skills, using different methods, and by reporting scores on each of these tests, not simply more but also better information is obtained about candidates, and as a result better decisions are made. If the decision made is for acceptance with some additional language programme, the test score information is available for diagnostic use. For a test such as M2, which is a direct performance test, and which according to experience reveals within-writer differences in some cases, extending the profiling to a more finely-tuned level not only aids reliability but is also a powerful tool for diagnosis and remediation. The potential of the instrument, as it now exists, for diagnosis and thence flexible placement is currently under investigation.

Appendix A

Original Assessment Scale ("OM")

BAND	BRIEF PERFORMANCE DESCRIPTION
9	**Expert Writer:** theme presented in a readable, intelligible, logical and interesting manner. Writes with complete accuracy and in the appropriate style. The reader is given a sense of mastery of the language and of the ability to handle the topic with complete competence.
8	**Very Good Writer:** theme presented clearly and logically, with accurate language forms and good style. Only very occasional inaccuracy or inappropriacy but which does not affect the communication. The reader can follow with no strain and will appreciate the argument expressed.
7	**Good Writer:** theme presented in a well-ordered, intelligible manner with well-structured and relevant supporting detail. Generally accurate in language and appropriate in style, but occasional lapses can affect the communication on first reading. The reader has, however, the impression of a functionally efficient writer.
6	**Competent Writer:** theme presented fairly logically and intelligibly. Reasonably accurate use of the language system. May have inaccuracies of style and presentation but showing an adequate functional competence. Can be read with only occasional strain put on comprehension.
5	**Modest Writer:** theme can be followed, but logical presentation may be broken and lack clarity or consistency. Several inaccuracies and style not always appropriate to presentation. May lack interest or variety, but the basic message is presented. The reader will have to strain on occasion to comprehend meaning.
4	**Marginal Writer:** theme can be followed with effort, and closer reading reveals lack of logical structure, clarity and consistency. Inaccurate vocabulary and sentence use coupled with inadequate connectors and cohesive features. Elements of information required may be omitted, repeated or inappropriately expressed. The reader has general difficulty in working out the message, though can eventually do so.
3	**Extremely Limited Writer:** elements of the information required are provided, but the presentation lacks any coherence. Uses over-simple sentence structure and impoverished vocabulary with continual errors and inappropriateness. Below level of functional competence though the reader may work out the general message.

2 **Intermittent Writer:** elements of the information required not provided, although a general meaning comes through intermittently. Either copies or produces strings of words. No real communication, although the reader may work out the general message.

1 **Unassessable Writer:** to be used for the true non-writer where no assessable strings of continuous English writing have been produced. OR: answer has been lifted 'en bloc' from Source Booklet, or a clearly irrelevant stock answer has been reproduced.

0 Should only be used where a candidate did not attend or attempt this part of the test in any way (i.e. did not submit an answer paper with his/her name and candidate number written on).

The British Council 1984

Appendix B

Profile Method ("PM1")

Profile Grid

Communicative Quality	9	8	7	6	5	4	3	2	1
Organization	9	8	7	6	5	4	3	2	1
Argumentation	9	8	7	6	5	4	3	2	1
Linguistic Appropriacy	9	8	7	6	5	4	3	2	1
Linguistic Accuracy	9	8	7	6	5	4	3	2	1

The British Council 1984

Appendix C

Revised Assessment Scale and the Global Method ("RM")

9 The writing displays an ability to communicate in a way which gives the reader full satisfaction. It displays a completely logical organizational structure which enables the message to be followed effortlessly. Relevant arguments are presented in an interesting way, with main ideas prominently and clearly stated, with completely effective supporting material; arguments are effectively related to the writer's experience or views. There are no errors of vocabulary, spelling, punctuation or grammar and the writing shows an ability to manipulate the linguistic systems with complete appropriacy.

8 The writing displays an ability to communicate without causing the reader any difficulties. It displays a logical organizational structure which enables the message to be followed easily. Relevant arguments are presented in an interesting way, with main ideas highlighted, effective supporting material and they are well related to the writer's own experience or views. There are no significant errors of vocabulary, spelling, punctuation or grammar and the writing reveals an ability to manipulate the linguistic systems appropriately.

7 The writing displays an ability to communicate with few difficulties for the reader. It displays good organizational structure which enables the message to be followed without much effort. Arguments are well presented with relevant supporting material and an attempt to relate them to the writer's experience or views. The reader is aware of but not troubled by occasional minor errors of vocabulary, spelling, punctuation or grammar, and/or some limitations to the writer's ability to manipulate the linguistic systems appropriately.

6 The writing displays an ability to communicate although there is occasional strain for the reader. It is organized well enough for the message to be followed throughout. Arguments are presented but it may be difficult for the reader to distinguish main ideas from supporting material; main ideas may not be supported; their relevance may be dubious; arguments may not be related to the writer's experience or views. The reader is aware of errors of vocabulary, spelling, punctuation or grammar, and/or limited ability to manipulate the linguistic systems appropriately, but these intrude only occasionally.

5 The writing displays an ability to communicate although there is often strain for the reader. It is organized well enough for the message to be followed most of the time. Arguments are presented but may lack relevance, clarity, consistency or support; they may not be related to the writer's experience or views. The reader is aware of errors of vocabulary, spelling, punctuation or grammar which

intrude frequently, and of limited ability to manipulate the linguistic systems appropriately.

4 The writing displays a limited ability to communicate which puts strain on the reader throughout. It lacks a clear organizational structure and the message is difficult to follow. Arguments are inadequately presented and supported; they may be irrelevant; if the writer's experience or views are presented their relevance may be difficult to see. The control of vocabulary, spelling, punctuation and grammar is inadequate, and the writer displays inability to manipulate the linguistic systems appropriately, causing severe strain for the reader.

3 The writing does not display an ability to communicate although meaning comes through spasmodically. The reader cannot find any organizational structure and cannot follow a message. Some elements of information are present but the reader is not provided with an argument, or the argument is mainly irrelevant. The reader is primarily aware of gross inadequacies of vocabulary, spelling, punctuation and grammar; the writer seems to have no sense of linguistic appropriacy, although there is evidence of sentence structure.

2 The writing displays no ability to communicate. No organizational structure or message is recognizable. A meaning comes through occasionally but it is not relevant. There is no evidence of control of vocabulary, spelling, punctuation or grammar, and no sense of linguistic appropriacy.

1 A true non-writer who has not produced any assessable strings of English writing. An answer which is wholly or almost wholly copied from the input text or task is in this category.

0 Should only be used where a candidate did not attend or attempt this part of the test in any way (i.e., did not submit an answer paper with his/her name and candidate number written on).

The British Council 1986

Appendix D

New Profile Scale and Profile Method 2 ("PM2")

	Communicative Quality	Organization	Argumentation	Linguistic Accuracy	Linguistic Appropriacy
9	The writing displays an ability to communicate in a way which gives the reader full satisfaction.	The writing displays a completely logical organizational structure which enables the message to be followed effortlessly.	Relevant arguments are presented in an interesting way, with main ideas prominently and clearly stated, with completely effective supporting material; arguments are effectively related to the writer's experience or views.	The reader sees no errors of vocabulary, spelling, punctuation or grammar.	There is an ability to manipulate the linguistic systems with complete appropriacy.
8	The writing displays an ability to communicate without causing the reader any difficulties.	The writing displays a logical organizational structure which enables the message to be followed easily.	Relevant arguments are presented in an interesting way, with main ideas highlighted, effective supporting material and they are well related to the writer's own experience or views.	The reader sees no significant errors of vocabulary, spelling, punctuation or grammar.	There is an ability to manipulate the linguistic systems appropriately.
7	The writing displays an ability to communicate with few difficulties for the reader.	The writing displays good organizational structure which enables the message to be followed without such effort.	Arguments are well presented with relevant supporting material and an attempt to relate them to the writer's experience or views.	The reader is aware of but not troubled by occasional minor errors of vocabulary, spelling, punctuation or grammar.	There are minor limitations to the ability to manipulate to linguistic systems appropriately which do not intrude on the reader.
6	The writing displays an ability to communicate although there is occasional strain for the reader.	The writing is organized well enough for the message to be followed throughout.	Arguments are presented but it may be difficult for the reader to distinguish main ideas from supporting material; main ideas may not be supported; their relevance may be dubious; arguments may not be related to the writer's experience or views.	The reader is aware of errors of vocabulary, spelling, punctuation or grammar, but these occasionally.	There is limited ability to manipulate the linguistic systems appropriately, but this intrudes only occasionally.

	Communicative Quality	Organization	Argumentation	Linguistic Accuracy	Linguistic Appropriacy
5	The writing displays an ability to communicate although there is often strain for the reader.	The writing is organized well enough for the message to be followed most of the time.	Arguments are presented but may lack relevance, clarity, consistency or support; they may not be related to the writer's experience or views.	The reader is aware of errors of vocabulary, spelling, punctuation or grammar which intrude frequently.	There is limited ability to manipulate the linguistic systems appropriately, which intrudes frequently.
4	The writing displays a limited ability to communicate which puts strain on the reader throughout.	The writing lacks a clear organizational structure and the message is difficult to follow.	Arguments are inadequately presented and supported; they may be irrelevant; if the writer's experience or views are presented their relevance may be difficult to see.	The reader finds the control of vocabulary, spelling, punctuation and grammar inadequate.	There is inability to manipulate the linguistic systems appropriately, which causes severe strain for the reader.
3	The writing does not display an ability to communicate although meaning comes through spasmodically.	The writing has no discernible organizational structure and a message cannot be followed.	Some elements of information are present but the reader is not provided with an argument, or the argument is mainly irrelevant.	The reader is primarily aware of gross inadequacies of vocabulary, spelling, punctuation and grammar.	There is little or no sense of linguistic appropriacy, although there is evidence of sentence structure.
2	The writing displays no ability to communicate.	No organizational structure or message is recognizable.	A meaning comes through occasionally but it is not relevant.	The reader sees no evidence of control of vocabulary, spelling, punctuation or grammar.	There is no sense of linguistic appropriacy.
1	A true non-writer who has not produced any assessable strings of English writing. An answer which is wholly or almost wholly copied from the input text or task is in this category.				

INVESTIGATIONS INTO THE APPROPRIATE USE OF VARIOUS ITEM TYPES FOR TESTING SPECIFIC LEVELS OF READING PROFICIENCY

Sandra McIntyre
Defense Language Institute

The purpose of this paper is to report results of preliminary investigations carried out by the Defense Language Institute concerning the appropriate use of item types for testing reading at various levels of proficiency. The investigations made extensive use of the language proficiency scale associated with the Interagency Language Roundtable (ILR) oral interview system and were specifically concerned with item types found in the standardized group-administered Defense Language Proficiency Tests.

Background

The ILR Oral Interview and the DLPT III

The language proficiency scale of the Interagency Language Roundtable is an eleven-point scale which defines levels of language proficiency for each of four skills: Speaking, Listening, Reading, and Writing.

The scale originates from the Interagency Language Roundtable Oral Interview system, which has been used by various government agencies as an official system for evaluating language proficiency for more than 30 years. During that time the proficiency levels have been carefully defined and they now serve as the basis for group-administered tests as well as for the face-to-face oral interviews. Some of the most widely used group-administered language proficiency tests are the Defense Language Proficiency Tests (DLPTs).

The DLPT III Reading Test

The Lower Range DLPT III Reading Test is a group-administered paper-and-pencil test which is machine scorable. It has five to seven parts, 100 to 120 items, and takes three hours to administer. Examinees are not permitted to use lexical aids or take notes. Item types that may be found in the DLPT III include the identification of signs, the identification of underlined information, factual and inferential comprehension questions, true-false-not addressed questions, and a modified version of the cloze.

The DLPT IIIs are based on the proficiency scale developed, refined and approved by the Interagency Language Roundtable. They are designed, written and validated according to the Interagency Language Roundtable Skill Level Descriptions. The validation process yields information regarding the correlation between proficiency levels of the examinees, as measured by the face-to-face assessment of reading during the oral interview, and the total scores of the examinees on the reading portion of the DLPT III. The validation data also provide an opportunity to examine the relationship between the proficiency levels of the examinees and the performance of the examinees on each part of the test.

Each part of the test uses an item format intended to measure a specific level of proficiency. Therefore, by examining the relationship between the proficiency levels of the examinees (as defined by the ILR) and their performance on each part of the test, information can be gathered concerning the appropriateness of using the item formats for testing specific levels of proficiency. The relationship can be examined by graphing the performance of the examinees at each level of proficiency.

For example, Appendix A shows the performance of examinees on the Reading Test of the Spanish DLPT III. The reading proficiency scores of the examinees, as determined by the reading portion of the face-to-face Interagency Oral Proficiency Interview, are shown along the x-axis. The percent of the items answered correctly on the DLPT III reading test is shown on the y-axis. Each of the curves represents a part of the test. For example, Part I, which consists of the "sign" format is represented by the curve at the top of the graph.

The intercept and the relative position of the curve in relationship to the other curves provides information concerning the difficulty of the subtest. In this case, the curve for Part I seems to indicate that it is easier than the other parts. Examinees at each of the six levels (0 + through 3) answered a higher percent of the items on Part I correctly than they did on the other parts. The shape of the curve provides information concerning the ability of the subtest to discriminate between the examinees of various levels of proficiency. The curve for Part I shows a sharp incline between the level 0 + and level 1 examinees. On the average, level 0 + examinees responded correctly to 50% of Part I items, whereas level 1 + examinees responded correctly to 75% of Part I items. The slope suggests that the item format in Part I discriminates well between examinees at levels 0 + to 1. Thus, by examining the curves for each of the item formats used in several DLPT IIIs for different languages, information has been gathered concerning the appropriate use of specific item formats for testing at the various levels of proficiency.

Formats

Signs

Part I of the DLPT III uses signs in the language being tested followed by four options in English. For example:

Exit (written in the target Language)

This sign says:

A. stop
B. exit
C. slow
D. caution

The sign format used in Part I was designed to test at the lowest levels of proficiency. The data graphed on curves generally support the intended use of the format. As previously mentioned, the results from the Spanish DLPT III (see

Appendix A) demonstrate that Part I was the easiest part for the examinees and that it discriminated best between levels 0 + and 1. The data obtained for the Korean, Chinese and French DLPT IIIs show a similar pattern of results for the sign format. (See Appendices B, C and D.) In each case Part I is relatively easy in comparison to the sections of the test which were designed to test beyond level 1 +. In addition, the slopes for Part I indicate that the format served to discriminate between examinees at levels 0 + to 1. Surprisingly, the curve of the Chinese DLPT III indicates a rise in the slope between the Part I scores of level 1 + and level 2 examinees which is larger than the rise between the level 0 + and 1 examinees. This slight aberration may be attributed to the small N of the 0 + group of students. It may also indicate the particular nature of Chinese signs which tend to represent relatively condensed language and low frequency character patterns. The overall results of the sign format, however, indicate that the format seems to be working effectively to discriminate among examinees at the lower end of the proficiency scale.

Information Identification

Appendix E provides an example of a second item type: the identification of underlined information. In the DLPT III the paragraph is in the language of the test and the options are in English. By varying the difficulty level of the paragraph, the underlined items and the distractors, the overall difficulty level of the format can be altered. The format has been used in the DLPT III to test reading proficiency at levels 1 + and 2 +. Appendix F is an example of information identification a level 2/2 +.

Both the Korean and the Chinese DLPT IIIs contain information identification items at levels 1 + and 2 + (see Appendices B and C). The relative positions of the curves suggest that the part containing the level 2 + items is more difficult than the part containing the level 1 + items. In both the Korean and the Chinese DLPT IIIs the level 2 + section continued to discriminate up through level 3, whereas the level 1 + section failed to discriminate between examinees at level 2 + and 3.

In summary, the information identification format has been used at levels 1 + and 2 +. Results suggest that some formats can be used for testing more than one level by changing the difficulty of both the stimulus and the options.

Comprehension Questions

Another format which has been used at different levels consists of a passage, in the target language, followed by comprehension questions in English (see Appendix G). The DLPT III reading tests have used this format with level 1 + passages and with level 2 + to 3 passages. For the level 2 + /3 section of the test, the difficulty of both the passages and the questions is increased. The questions used for testing at level 1 + ask for concrete factual information. Those used for testing at level 2 + /3 tend to be much more inferential in nature. Although developing multiple choice comprehension questions at the 2 + /3 level requires more effort than developing questions at the 1 + level, the format does seem to work as well at the higher level as at the lower level.

Part II of the Korean DLPT III (see Appendix B) consists of short passages, such as common announcements, followed by factual questions. The difficulty level of Part II is comparable to that of Part I. The curve for Part II also tends to follow approximately the same slope as Part I, though Part II discriminates more powerfully than Part I between levels 1 and 1 + .

Part IV of the Chinese DLPT III (see Appendix C) consists of relatively long passages of a more abstract, theoretical nature, followed by inferential multiple choice questions. The relative placement of the curve suggests that the examinees found it difficult. In addition, Part IV discriminates well among the higher level examinees. In conclusion, the multiple choice comprehension questions format has been used successfully to test higher levels of proficiency (2 + /3) as well as the lower levels of proficiency.

True-False-Not Addressed

Another format which has been used to test reading at the abstract/inferential levels (levels 3 through 3 +) is the True-False-Not Addressed format. An example can be found in Appendix H.

The DLPT III contains authentic passages in the target language followed by statements in English. The examinees are instructed to determine, on the basis of the information given in the passage, whether a statement is true, false, or not addressed. A statement is true if it can be confirmed on the basis of the information given in the passage. It is false if it is contradicted by information given in the passage, and it is not addressed if there is insufficient information in the passage to either confirm or deny the statement. Although this format may appear to be fairly easy to construct, careful attention must be given to determining and confirming through field trial that each statement is, in fact, clearly true, false, or not addressed. This format, however, has been found to work well in testing the ability to understand ideas and inferences at level 3/3 + .

Part V of the Chinese DLPT III reading test is more difficult than Part IV for the students who are at the level 2 to 3 range of proficiency. (See Appendix C.) The reversal of the curves for level 0 + and 1 + examinees may be attributed to the fact that the chance of correctly answering an item in Part V is higher than the chance of correctly answering an item in Part IV. (Part V has only three options from which to select, whereas Part IV has four options.) The steep and steady incline in the slope for the Part V curve among the level 1 + through level 3 examinees illustrates that the True-False-Not Addressed format is useful at this level.

Cloze

The traditional cloze format, in which every seventh word is randomly chosen for deletion from the text and the examinees are asked to use their intuition to predict the best word to fill in the blank, has been significantly modified for use in the DLPT III. Because the DLPT III is a machine-scorable test, a list of options must be supplied. In addition, the lower level cloze passages are prefaced by an English paraphrase of the passage. Appendix I is an example of the format.

The results of the cloze format in the Spanish DLPT III (see Appendix A) are fairly representative of the results found for other languages as well. In general, the cloze passages tend to be quite difficult. The relative positions of the curves show a large increase in difficulty level between Part I (level 0 + and 1) and Part 2 (level 2 cloze), and the Part IV underlined information (level 2 +) is easier than the level 2 cloze.

Verbal comments given by the examinees also indicated that the cloze format was generating a high level of frustration. Much of the difficulty may have been due to a test-administration limitation in which the examinees were not permitted to mark in the booklets. Therefore, there was an additional load on short-term memory from the requirement of having to remember which options had already been selected.

The cloze format, however, does seem almost unique in its ability to discriminate with a very high reliability and power across a wide range of reading proficiency levels. Each of the curves for the cloze format shows a consistent, steep incline in the slope from level 1 through level 3. Statistically, the cloze works well, and it is relatively easy to construct in comparison to the multiple choice and true-false-not addressed formats.

It has been suggested that further investigations into the difficulty level of the cloze format might include using the same passages which were used for the cloze in an alternate format (such as true-false-not addressed, multiple choice comprehension questions, or traditional cloze formats). Such investigations would help to pull apart the interaction between the effects of item format and difficulty level of the passage.

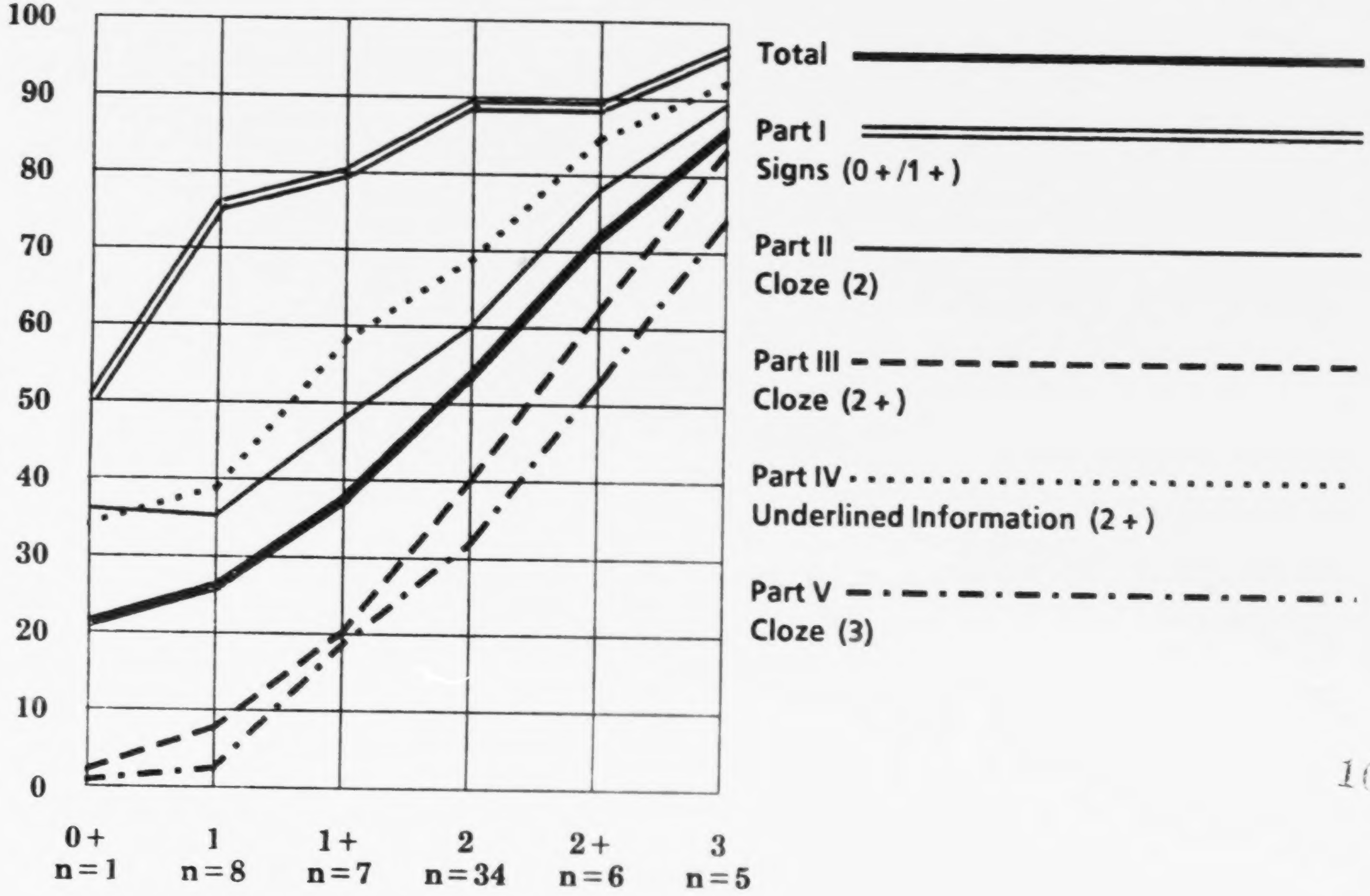
Conclusion

At DLI we have incorporated the results of these investigations to make the DLPT III reading tests even more effective as a measure of reading proficiency. The first DLPT IIIs, for example, employed the design exemplified by the Spanish test, and although the test had high reliability and validity, it generated complaints from examinees. As illustrated in the graph, the jump in difficulty between Part I and Part II was quite large. The general frustration level led examinees to question the face validity and suitability of the task. As a result, a new design was developed for the DLPT III reading test.

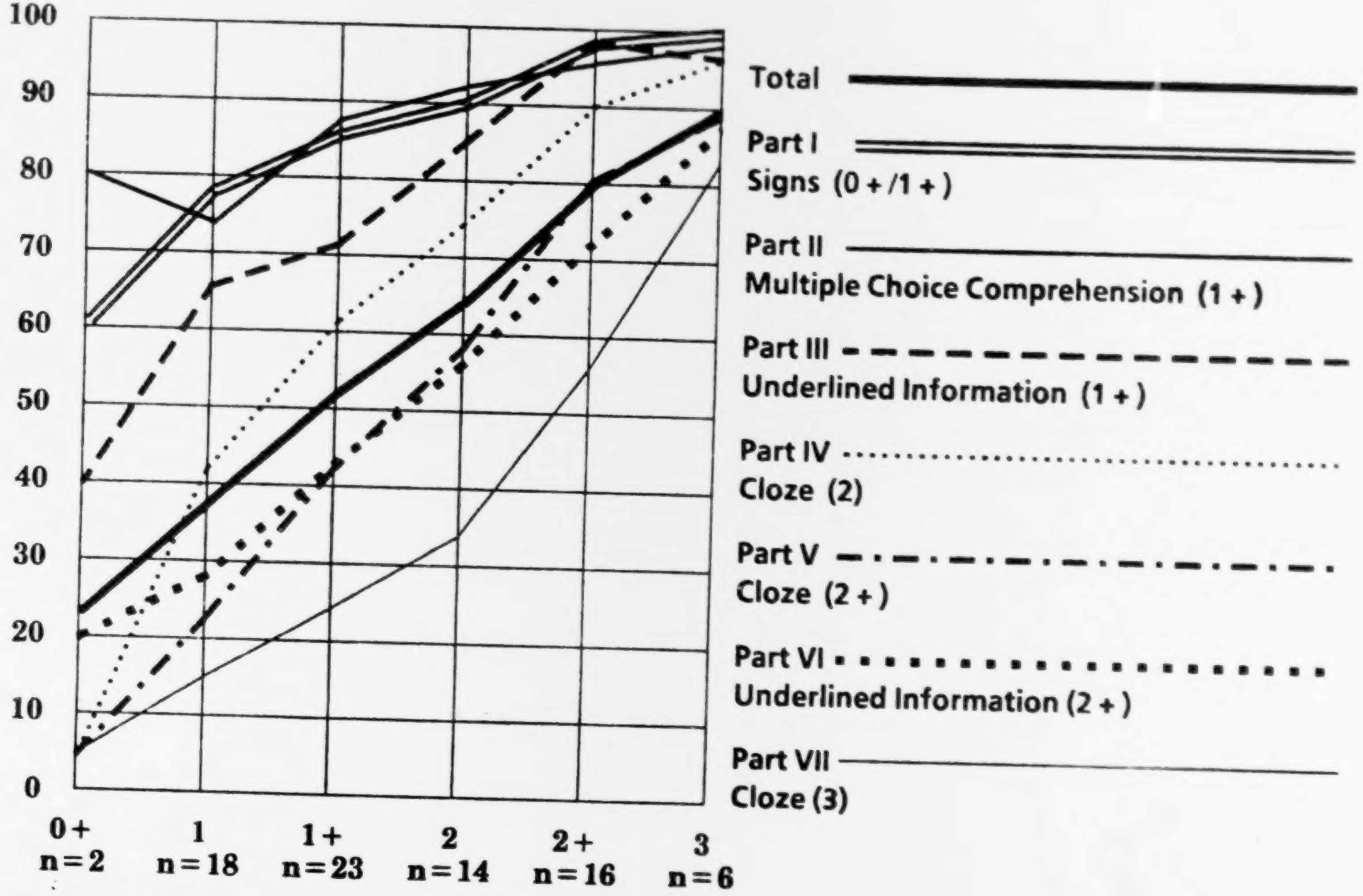
The new design incorporated several changes. The information format was moved to an earlier part of the test, and the cloze was replaced by inferential comprehension questions and true-false-not addressed items. Both of these formats had previously been used successfully in the listening portion of the DLPT III. The progression of the difficulty levels for the parts of the new DLPT III reading design has been satisfactory, and the reliability coefficients have remained strong.

In conclusion, it should be noted that the ILR scale has been very useful not only in the initial development stages of the DLPT III, but also in its later validation and revision. The ILR scale has played a major role in optimizing the validity of the DLPT III tests. Not only are the test designs based upon the ILR definition, which give them a certain inherent face validity, but the results of the test can also be examined by using the ILR scale to determine criterion-related validity. The oral interview, which can be used to assess three skills face-to-face, serves as an external criterion with which a correlation can be examined between total performance on the DLPT III and on the Oral Interview. In addition, the parts of the test can be examined based on Oral Interview performance. Without an integrated system based on a well defined scale, it would be a more difficult task to interpret the results of similar statistical analyses. Basing a test on a well-defined scale facilitates the design, development and validation of the test.

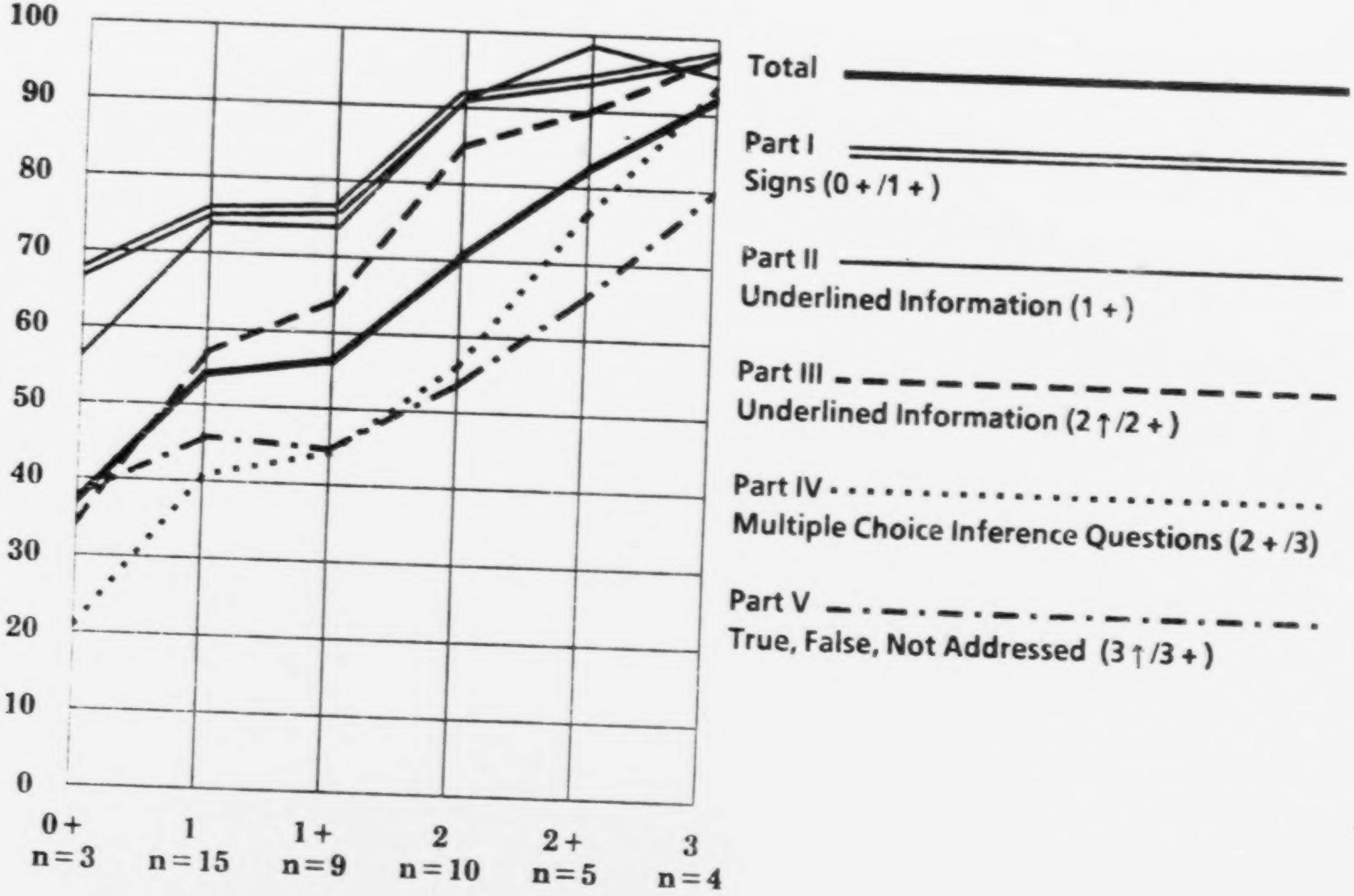
Spanish DLPT III
Reading



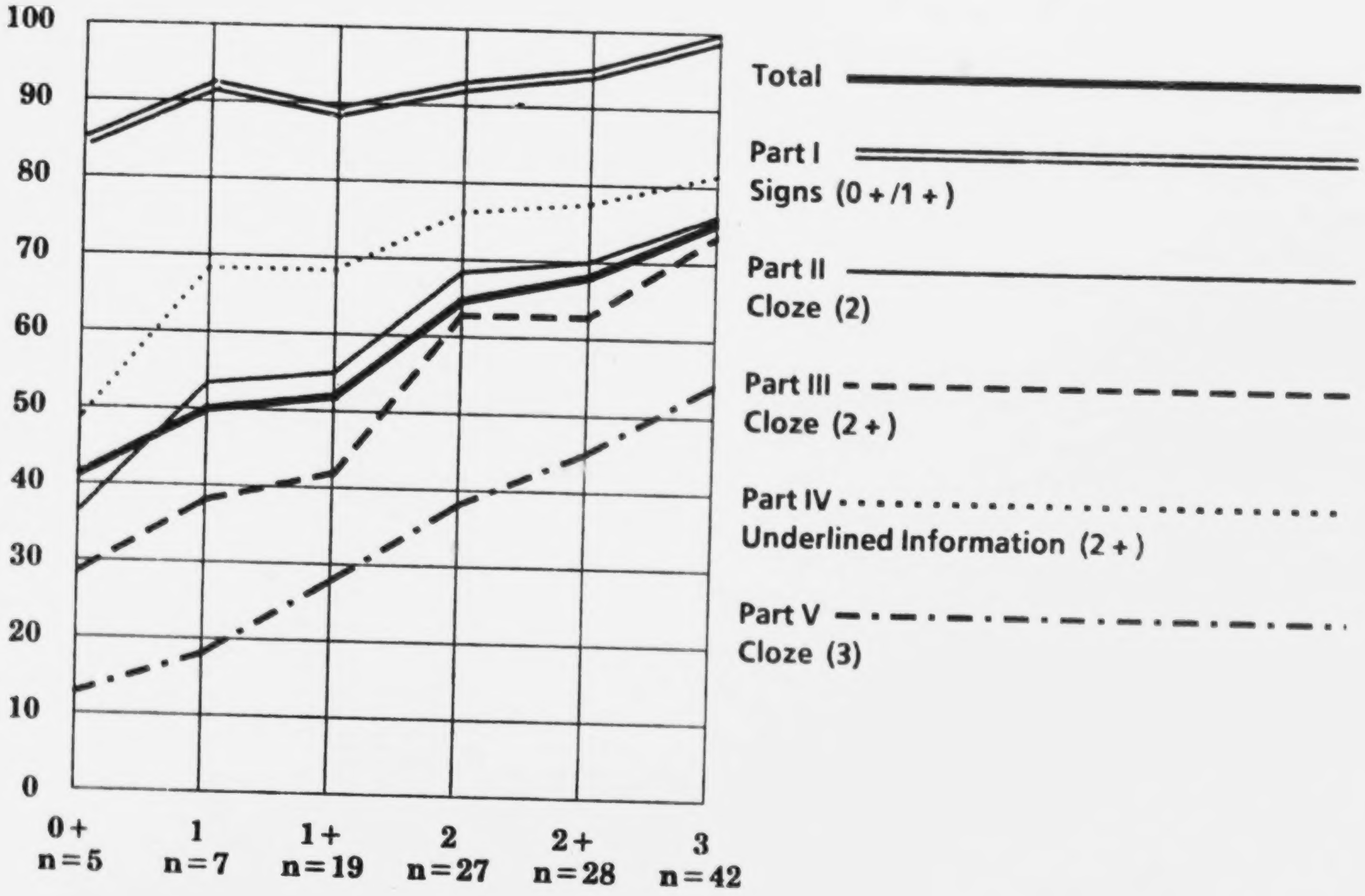
Korean DLPT III
Reading



Chinese Mandarin DLPT III
Reading



French DLPT III
Reading



Appendix E

Information Identification, Level 1 +

Human Relations Meeting Set

The Seaside Human Relations Commission will hold a special $\underset{1}{\underline{\text{meeting}}}$ today $\underset{2}{\underline{\text{to plan activities}}}$ and projects $\underset{3}{\underline{\text{for the year}}}$. The meeting $\underset{4}{\underline{\text{will be held}}}$ at 5:30 p.m. in the $\underset{5}{\underline{\text{City Hall}}}$ Conference room.

1.
   - A. meeting
   - B. dance
   - C. dinner
   - D. show

2.
   - A. to plan activities
   - B. planned to be active
   - C. plan to act it out
   - D. had planned to act on it

3.
   - A. for the year
   - B. after one year
   - C. a year ago
   - D. took a year

4.
   - A. will be held
   - B. was held
   - C. will be given
   - D. was given

5.
   - A. City Hall
   - B. Community Center
   - C. Concert Hall
   - D. Public Library

Appendix F

Information Identification, Level 2/2 +

Police Arrest Reeling Horseman For Drunken Driving After Chase

FREEMONT (AP)--A man who rode his horse home from a bar was arrested for[1] investigation of drunken driving[2] after a policeman saw him nearly fall out of the saddle[3].

John Charles Black was released on $1,500 bail[4] after his arrest early Monday. He was also accused of resisting arrest[5] because the horse allegedly[6] galloped off when the pursuing[7] police officer turned on his siren[8].

His wife, Tammi, said she and her husband often ride home on horseback from bars. "After all, you can't hurt yourself[9] on a horse and the horse knows its way back," she said. "It's the way to go if you're going to go out drinking.

"The horse watches out for cars ... It gives you time to sober up[10] between the bar and home."

But Freemont police said a horse falls under[11] the state Vehicle Code[12] definition[13] of a vehicle: anything that can be "propelled[14], moved or drawn upon a highway[15]."

1.
A. arrested for
B. questioned about
C. wanted for
D. suspected of

2.
A. drunken driving
B. reckless driving
C. improper conduct
D. speeding

3.
A. fall out of the saddle
B. collapse in his car
C. fall into this sad state
D. speed into the divider

4.
A. released on $1,500 bail
B. paid a fine of $1,500
C. held on a $1,500 bail
D. paid $1,500 in damages

5.
A. accused of resisting arrest
B. requested to stay still
C. warned to slow down
D. wanted to take a rest

6.
A. allegedly
B. quickly
C. willingly
D. suddenly

7.
A. pursuing
B. suspecting
C. determined
D. unsuspecting

8.
A. turned on his siren
B. turned up in his path
C. turned in his license
D. turned up his bullhorn

9.
A. hurt yourself
B. hurt others
C. be hurt by others
D. be hurting others

10.
   A. sober up
   B. unwind
   C. exercise
   D. slow down

11.
   A. falls under
   B. falls down
   C. slips below
   D. slips up

12.
   A. Vehicle Code
   B. Driver's License
   C. Code of Conduct
   D. Driver's Manual

13.
   A. definition
   B. law
   C. expiration
   D. explanation

14.
   A. propelled
   B. flown
   C. sped
   D. ridden

15.
   A. highway
   B. runway
   C. street
   D. alley

Appendix G

Comprehension Questions, Level 2 + /3

A Soviet scheme to divert the flow of two great Siberian revers southward to irrigate new agricultural lands "may have significant repercussions for the weather patterns of at least a hemisphere, if not the whole globe," writes physicist John Gibbin in the liberal Guardian of London (July 16).

Soviet planners are determined to reverse much of the flow of the Pechora River and the Ynisei and Ob Rivers. This will provide vital irrigation to millions of acres in Kazakhstan, a region opened to agriculture only in the past thirty years, and triple the grain production in the huge area. Dr. Gibbon says, "as far as it is possible to tell, the Soviet planners have dismissed the possibility that the side effects might do more harm than the irrigation does good."

The diversion will reduce by 20 percent the fresh water flow into the Arctic Ocean, "The ice-covered Arctic is the key factor in establishing the climatic patterns of the Northern Hemisphere." The diversion will allow the warmer salt water to rise to the surface and melt the ice. "By reducing the fresh water flow and causing the ice to break up, the irrigation schemes may directly reduce the rainfall in exactly the regions they are designed to help."

1. According to the article, the two great Siberian rivers

   A. are a matter of future international concern.
   B. are gradually yielding less water for irrigation.
   C. are threatened by changes in global weather patterns.
   D. currently flow southward into the grain producing region.

2. The plan will

   A. increase the water supply from the source into the rivers.
   B. revitalize ancient agricultural areas.
   C. increase the production in a relatively new agricultural region.
   D. forestall the effects of the expected weather pattern changes.

3. Dr. Gibbon feels that

   A. the Soviets have not recognized the possible repercussions.
   B. the plan will help the Soviets at the expense of other nations.
   C. the Soviets are wrong in expecting major weather pattern changes.
   D. the benefits of the plan outweigh the possible side effects.

4. Twenty percent of the fresh water

   A. will no longer flow into the Arctic Ocean.
   B. will come from melted ice from the Arctic Ocean.
   C. is expected to be lost into the ocean in the next few years.
   D. is expected to evaporate during the coming year.

5. According to the article

   A. the Soviets' plan may backfire.
   B. the plan is admirable and demonstrates foresight.
   C. weather patterns can't be controlled by man.
   D. agriculture is at the mercy of natural weather conditions.

Appendix H

True-False-Not Addressed, Level 3

BELGRADE, Yugoslavia--The Yugoslav leadership has warned the press to follow the party line and resist the arguments of opponents of the country's unorthodox self-management system of communism.

A statement by the nine-man collective presidency last week said "forces of international reaction" are bringing pressure on Yugoslavia in an effort to destabilize it.

Their object is "to bring in doubt and spread despondency and disbelief in Yugoslavia's capacity to overcome its problems," the statement maintained.

Yugoslavia's grave economic problems include debts to the West of $20 billion, high inflation, an acute shortage of hard currency, stagnating industrial production and falling living standards.

The statement said hostile forces, which it did not identify, are trying to penetrate Communist institutions and to influence the media. Media editorial boards are giving a distorted picture of Yugoslav reality, it said.

The statement did not specify which articles are in disfavor.

Questions:

1. The Yugoslav leadership supports an unorthodox system of Communism. (T)
2. The leadership attributes the unrest to international factors rather than to internal economic problems. (T)
3. The reactionary forces claim that a crisis is inevitable if the $20 billion national debt continues to grow. (NA)
4. The information disseminated by the media continues to reflect the ideology of the leadership. (F)
5. The leadership warned that authors of reactionary articles face the threat of arrest and imprisonment. (NA)

Appendix I

Paraphrase **Cloze, Level 2**

On Thursday a 21 year old man identified as Michael W. Phillips from Seaside was hurt in an accident involving a motorcycle and a car. The driver of the car was Ho Jong Lee, 45, of Milpitas.

The accident occured in Monterey shortly before 3 p.m. Phillips will be treated before being released from Community Hospital.

Passage

Motorcyclist Injured In Collision

A seaside man was ___★___ Thursday ___●___ when ___●___
1 2 3
motorcycle ___★___ with a car on Garden Road in Monterey.
4
___●___ identified the injured ___●___ as Michael W. Phillips, 21
5 6
of 1148 Buena Street.

He ___■___ injured in ___★___ with a ___●___ driven ___■___
7 8 9 10
Ha Jong Lee, 45, of Milpitas, according to ___■___ police ___★___ .
11 12

Phillips was ___★___ to Community hospital, ___●___ a spokesman
13 14
___★___ it was expected ___■___ he would ___■___ released ___●___
15 16 17 18
treatment.

The accident, which occurred ___●___ before 3 p.m., was ___■___
19 20
Garden Road north of Olmstead Road.

★

A. reported
B. report
C. taken
D. injured
E. collided
F. collision
G. said

●

A. where
B. when
C. after
D. just
E. afternoon
F. policeman
G. police
H. car
I. motorcyclist
J. his

■

A. was
B. be
C. on
D. by
E. a
F. that

GRADING READING PASSAGES ACCORDING TO THE ACTFL/ETS/ILR READING PROFICIENCY STANDARD: CAN IT BE LEARNED?[1]

Dale L. Lange
University of Minnesota

Pardee Lowe, Jr.
Central Intelligence Agency

Introduction

In its American Council on the Teaching of Foreign Languages (ACTFL)/Educational Testing Service (ETS) incarnation (1982, rev. 1986), the Interagency Language Roundtable (ILR) Oral Proficiency Scale and its accompanying levels have been successfully transferred to academe. In contrast, the ILR reading proficiency scale, even in its ACTFL/ETS form, has been less well received. Those attempting to apply it for the first time consistently comment that the reading scale seems harder to grasp than the oral one. Moreover, they understand the rationale for using the reading scale less fully. And more importantly, they question both whether reading proficiency test performances in academe can be rated accurately according to the scale and particularly whether the passages, the comprehension of which forms the basis for ratings, can be properly graded for level.

Purpose of the Study

This study investigates the execution of two tasks pivotal to assigning levels to reading passages: ranking and rating. These tasks were specifically chosen to demonstrate that the ACTFL/ETS and ILR (hereafter referred to as AEI) reading scales provide a meaningful basis for rank-ordering and assigning levels of reading competence. Ranking, the easier of the two tasks, is generally not formally separated from rating as a task. It was chosen, however, to demonstrate to the participants of the item writing workshop that general degrees of difficulty are inherent in passages, a view that contrasts with one regarding each passage as unique and unrankable. If it can be demonstrated that passages possess varying degrees of difficulty, it can be seen that rating by generalized categories, such as the ACTFL/ETS Guidelines, becomes a logical next step. Rating was selected to show that a passage's level of difficulty could usually be matched to a verbal definition without the definition describing every aspect of a passage.

---

[1] We wish to thank Martha Herzog, John Lett, and Ray T. Clifford at the Defense Language Institute for reading an earlier draft of the paper. Their comments have been invaluable. To Martha Herzog and her colleagues at the Defense Language Institute also our gratitude for the excellent selection of texts and for providing insightful commentary on their use in face-to-face reading proficiency interviews.

The Context

The setting for this study was a proficiency test writing workshop at the University of Minnesota in the Summer of 1985 as part of the University's Foreign Language Project. The project focuses on a new language requirement for entrance to the College of Liberal Arts and for graduation with a Bachelor of Arts degree from the college. Arendt, Lange, and Wakefield (1986) have described the project in detail. The new language requirement demands functional competence in a second language in listening, reading, writing, and speaking as opposed to seat time. It is based upon the AEI proficiency statements.[2] (See the attached Guidelines for the generic descriptions.)

A series of steps in the form of three workshops was organized to develop the testing program for the new language requirement. In the first one, participants from the University language departments, the community colleges, private liberal arts colleges, the state university system, and public schools established the expected levels of proficiency in listening, reading, speaking, and writing for both an entrance standard and a graduation requirement. The chosen levels, which follow, can be interpreted with the Guidelines presented in the appendix:

	Entrance Standard	**Exit Requirement**
Listening	Intermediate Low	Intermediate High
Reading	Intermediate Low	Intermediate High
Speaking	Novice High	Intermediate Mid
Writing	Novice High	Intermediate Mid

A detailed statement of expectations for each of the modalities was also constructed. Called a "functional trisection," each statement contains descriptions of the content, functions, and accuracy to be demonstrated by students with each modality for both the entrance standard and graduation requirement.

The second workshop concentrated on the testing of the four language modalities. Participants used the functional trisection they developed as the basis for the discussion of test items and test types, their limitations, and test constraints. They examined multiple-choice and true false items, cloze tests, and such variables as time and facilities for the testing of 1500 students.

---

[2] ACTFL/ETS/ILR (AEI) designates those aspects common to the two scales, such as the AEI concept of "proficiency" (details in Lowe [1986]). The term is not applied to those aspects which differ, such as testing procedures. For example, FSI employs a reading interview while DLI uses a multiple-choice paper-and-pencil testing instrument. In such cases, the procedure is designated by the individual user.

In a third workshop, one week in length, participants wrote a bank of items for the potential entrance standard and graduation tests in listening, reading, speaking, and writing. It was in this latter context that the study took place.

Procedures

Preparation

Although they were already familiar with the AEI scales in all modalities from prior workshops, the first step for this third workshop group was the review of the level definitions for Reading Guidelines on the first day. Further, there was a general discussion of factors contributing to the levels from the grading of a sample of some 27 reading passages by the ILR Testing Committee. The discussion helped participants focus their comprehension of the system on the tasks. After this discussion, the initial ranking and rating of eleven sample ILR texts took place. There was subsequently no discussion of the texts until the two tasks of rating and ranking had been accomplished. Once the ranking and the assignment of ratings were complete, a lively discussion about passage levels and their difficulty arose. This experience served as the basis for the preparation of items for listening, reading, speaking, and writing for the remainder of the week. The discussions helped reinforce the guidelines as a standard for their work.

The Passages

The texts range in difficulty from 0 + /Novice High to Distinguished or Level 5 proficiency descriptions. Due to space limitations, here we will discuss only three representative texts.

Text One, "TV."

> The 0 + text, labeled "TV," is a picture of a man carrying a TV set. He appears to be entering a TV repair shop. In the window, there are several signs: TV, Service, Closed Wednesdays. The text contains isolated high frequency words, supported by considerable visual context.

Text Two, "Second Man Held."

> 2nd Man Held in Twins' Deaths
>
> A second suspect was arrested yesterday in the shooting deaths Friday of Richard F. and Ronald F. Grey, 27-year-old twins, Brockton County police reported.
>
> Samuel K. Cummings, 29, of no fixed address, surrendered at 4 p.m., police said. He was held without bond in the county jail on two first-degree murder charges.
>
> The bodies of the Grey brothers, who lived on Freeville Road SE, were found in separate locations in Upper Boonsboro about 10 miles from the place where they were probably shot -- the rear of the '81' Club in Hampton Heights. Early Sunday, police charged Ralph P. Lucas, 26, of Hampton Heights with two counts of murder.

This is a text of Level 2 difficulty. It is basically factual, but the writer makes a number of assumptions about what readers will understand. For example, the "second suspect" is mentioned before indicating there was a first suspect. And "...bodies...were found in separate locations" is confusing because the victims are twins who presumably lived at the same address. Moreover, the reporter assumes the reader possesses the background knowledge of the legal system and an understanding of language to register the tentativeness of his statements. The text is full of past tense forms. The sentence length increases as does the level of abstraction of language. It appears that compound and complex sentences could play a major role in the comprehension of this text.

Text Three, "When."

> When we look around us today, we see tremendous sums of public and private money poured into artistic and cultural activity at every level. We see a vast network of institutions serving a large and eager but often bewildered public. And, not least, we also see a great deal of unmistakable talent and imagination at work.
>
> Yet how directionless and stymied, how baffled in their purposes, most of this activity and talent seems. In fact, after viewing the art scene all these years, it is impossible for me not to ask: What's wrong here?
>
> Let me put it another way: Why is so much of our art so empty and mean-spirited? Why do so many vaunted reputations turn to ashes so quickly? Why doesn't all the talent, effort, and money produce more of quality and permanence? Why is so much of the criticism lavished upon our art so pusillanimous in confronting failures? And why are our values, tastes, and intellectual loyalties so threadbare? Plainly, there are many things missing in our cultural life. One of the most important of these missing elements, it seems to me, is a critical perspective that is at once serious, high-minded, and disinterested--capable of producing criticism of such integrity that it stands apart from the blizzards of publicity and the unacknowledged social scenarios that today dominate the arts and traduce their objectives.
>
> To put it more bluntly, what is urgently needed in our artistic and cultural life is criticism that asks hard questions, challenges reigning orthodoxies, speaks up for quality and upholds a sense of standards.

This is a Level 4 text. Readers must delve four paragraphs into the text before they discover the topic of the article, "a critical perspective" on the arts. Beginning "in medias res" is a major characteristic of Level 4 texts with their unpredictable chains of thought, the author plunging the reader into the author-made, author-controlled, author-described world. Precise language, synonymous words, phrases, idioms, and constructions mould the reader's perceptions of the author's goal(s). The reader queries also, reading not only between the lines, but reading beyond, forced by the author's unorthodox approach, provocative statements, and careful choice of words to enter a world he does not anticipate. The author achieves his goal when readers emerge somewhere other than where they started.

Tasks and Directions

In the study, eleven new English passages, carefully rank-ordered and graded for level by the ILR Testing Committee, were given blind and in random order to the twenty-five participants for both ranking and rating.

In the first task, ranking, the participants were asked to rank-order the passages for difficulty. There were four rounds to this task. Specifically, the participants received the following instructions:

A. Using your own experience, not those of any possible student test population, rank the passages from easiest to hardest.

B. Assign 1 to the easiest passage, 2 to the second easiest, and so on.

C. Assign 11 to the hardest passage, 10 to the next hardest, and so on.

D. Try to force a choice between any given pair of passages.

E. If two items tie, then write a justification for assigning the same rank to the two passages.

Round One

The purpose of this round was to acquaint the participants with the new passages and to obtain a preliminary overall ranking. Subjects were told to rank all eleven passages and enter their ranking on the Round One reporting sheet (Figure 1 below).

Figure 1

Form Listing of Texts in Random Order

First Name: ____________________ Last Name: ____________________________

Passage Name	**Level**	**Round** _______
Andres Restaurant	______	
Citizens Advisory Committee	______	
Frank O'Hara	______	
Motor Injured	______	
OK Mommy	______	
Second Man Held	______	
They Call This Progress	______	
Time To Reconsider	______	
TV	______	
When We Look Around	______	

(Throughout this article, the texts are referred to by the first word in their titles.)

Rounds Two Through Four

The "focus" rounds, Rounds Two through Four, were introduced to lend greater precision to the rankings within any subset of passages, since it was assumed that the participants could discriminate more readily between passages at either end of the continuum than between contiguous ones. Consequently in filling out the reporting sheet for each round, participants were asked to focus in Round Two on the passages they had ranked 1-4, in Round Three on the passages they had numbered 8-11, and in Round Four on the passages they had designated 5-7. They were also told to adjust the rankings of passages other than the ones specifically selected in the round if their focussing mandated a reordering. It was understood that participants might differ in which passages they ranked 1-4, 5-7, and 8-11. In Round Four, participants were instructed to furnish their final overall ranking. We had originally planned to repeat this first task on the last day of the workshop, but the ranking results from the first day suggested that repetition was unwarranted.

In the second task, rating, participants were instructed to assign levels to the passages according to the ACTFL/ETS Guidelines for reading. (See Appendix.) Their choice of levels for each passage was then compared to the levels assigned by the ILR Testing Committee.

To understand how rating was carried out, we need to discuss the relationship of the ACTFL/ETS guidelines to the ILR scale. The relationships of the two systems' reading scales has grown more intricate since the ACTFL/ETS version's last revision. (See Figure 2.)

Figure 2

Comparison of ACTFL/ETS and ILR Scales

ACTFL/ETS SCALE LEVELS	ILR SCALE LEVELS
Superior (S, 3-5)	5
	4 +
	4
	3 +
	3
Advanced Plus (ADV +)	2 +
Advanced (ADV)	2
Intermediate High (IH)	1 +
Intermediate Mid (IM)	
Intermediate Low (IL)	1
NOVICE HIGH (NH)	0 +
NOVICE MID (NM)	
NOVICE LOW (NL)	0
– ABSOLUTE ZERO	

Like the ACTFL/ETS oral scales, the ACTFL/ETS reading scales subdivide the lowest ILR Levels 0 and 1 and assign them verbal descriptions: Novice and Intermediate. Unlike the ACTFL/ETS oral scales, however, the revised ACTFL/ETS reading scales provide numerical base level descriptions (ILR 3, 4, and 5, but not their plus levels) for what in the oral scale corresponds to the omnibus designation, Superior. The interrelationship of the ACTFL/ETS guidelines to the ILR scales became important to the study because some of the ILR sample reading passages required further subdivision (3, 3+, 4, 4+, and 5). (The ACTFL/ETS scale does not include the "plus" level distinctions.) Due to this confusion in the first round, we report only overall results. In the second round, the ILR distinctions were made. The details are given below.

In the rating task, participants were told to match the passage and the linguistic behaviors needed to comprehend it to the Guideline's level descriptions. They were specifically directed to:

A. Match the passage to a single ACTFL/ETS level description.

B. Bracket the passage with a description either side of the one you originally chose. (The purpose of bracketing is to provide three definitions for comparison to ascertain which definition best describes the passage and the behaviors a reader would have to employ to understand the passage.)

C. Determine which of the three descriptions most accurately reflects the nature of the passage and its level of language.

D. If a description to either side of your original choice seems to fit better, bracket again.

E. Repeat the process until you have made your final determination.

Such a procedure generally permits re-rating according to the ILR system except at the higher levels, where the ILR system makes more distinctions than the ACTFL/ETS system, as depicted above.

Results and Discussion

Ranking

Round One revealed that, on average, the participants accurately ranked five of the eleven passages, if one uses the ILR Testing Committee designations as the accuracy criteria. This round identified the two anchor passages, "TV" at the lower end and "Frank" at the higher end. While the rankings of individual participants sometimes varied widely, the averaged group ratings indicated that participants readily identified the polarity of these two extreme passages.

Round Two, with its focus on the easiest passages revealed eleven correct assignments out of eleven. ILR experience suggests that while it is easier than rating, ranking also requires practice, and consequently, slight variations are permissible. In ILR work, passages are generally ranked and rated by committee, as were both the passages for illustration and those for the blind rating used in this study.

Round Three, with its concentration on the hardest passages, produced nine correct ratings out of eleven. There were misassignments of two passages, "OK" and "Poetry." Round Four, which concentrated on the passages ranked in the middle of the range and provided a final overall rating, also produced nine correct assignments out of eleven. Obviously, with full agreement between the ILR ranking and the participants' average overall ranking in Round Two, the study could have concluded at that point. However, the last two rounds introduced slight variation for four passages: "OK" "Poetry," "They" and "Time."

Results for Part One demonstrate that a consensual basis exists for regarding some passages as harder than others and for ranking the passages accordingly. At this juncture, no overt comparison to the ACTFL/ETS Guidelines took place. Table 1 displays the comparative ranking for each round. Obtaining complete agreement was not the study's major goal, but a high degree of agreement was desirable, and it was achieved.

Table 1

Consensual Ranking of Reading Passages

Passage	Round				UMinn Average	ILR Ranking	ILR Level
	1	2	3	4			
Andres	3	4	4	4	4	4	1/1 +
Citizens	5	5	5	5	5	5	1 +
Frank	11	11	11	11	11	11	5
Motor	6	6	6	6	6	6	2
OK	2	2	2.5	3	2.4	2	1
Poetry	4	3	2.5	2	2.6	3	1
Second	7	7	7	7	7	7	2 diff
They	9	8	8	9	8.3	9	3
Time	8	9	9	8	8.8	8	2 +
TV	1	1	1	1	1	1	0 +
When	10	10	10	10	10	10	4 low

(In round 1 all the texts were ranked. Rounds 2 and 3 dealt with the easiest and hardest texts respectively. In round 4 the midmost texts and then all texts were ranked. Twenty-four participants completed the training.)

Rating

The eleven level assignments in Round One revealed seven complete agreements, three within a plus level of the ILR, and one within a level and a plus of the ILR. An inherent problem exists in assigning levels to passages, as Child (1986) has indicated. The ILR system is designed to rate the processing the reader undertakes, not the product, as in an oral recall protocol in FSI's reading interview. In the FSI procedure, the test taker is given a target language passage to read silently and then is asked to produce a gist. The ILR descriptions, like the ACTFL/ETS Guidelines, are expressed in terms of what a non-native can consistently and sustainedly comprehend. Consequently, to assess the candidate's performance, test administrators must grade each passage so that they can determine whether the candidate understood at the requisite level. This accurate assigning of a level to reading passages obviously requires practice, as does accurate rating of oral interviews.

In Round Two, the participants, on average, assigned the same (or an equivalent ranking) as the ILR Testing Committee did for ten of the eleven passages. In this round, those respondents writing "S" (ACTFL/ETS Superior, which subsumes ILR levels 3, 3 + , 4, 4 + , and 5) were asked to further define that designation, according to the ILR scale. This additional step increased the rate of correct assignments from seven in Round One to ten in Round Two.

Accurate rating depends on internalization of the standard. The fact that the internalization of the scale at the level of the individual is less advanced is shown in this study by the mean score on individual rater/ranker performances, which ranged from a total 3 to 10 out of 11, the mean being 6.5. Passages assigned a split rating, e.g., IM/H, were counted as being appropriately rated if one of the ratings matched that assigned to the passage by the ILR Testing Committee. This method of scoring, requiring the exact original rating or "exact scoring" (ESC) masks the important fact that participants usually rated within a plus point of the ILR Testing Committee's rating.

In rating oral interviews, another method of scoring is applied: in any group of 10 interviews, it is expected that the majority will receive the same ratings as those assigned by experienced testers, with a few deviant scores, no more than two or three within a plus point of an experienced tester's rating. Applying the same approach to the performances of novice passage raters, participants in the present study, scoring 8 to 10, qualify as "proficient" passage raters. Ten participants out of 24 qualified in that range. The ILR Testing Committee, however, has long recognized the difficulty in assigning levels to passages and regularly encourages testers to rate passages in groups. One could also apply a less strict method of scoring, "Proximate Scoring" (PSC), which recognizes that rater trainees are beginning to internalize the standard when they are within a plus point of the original ILR rating. Such scoring suggests that many individuals are indeed achieving internalization of the standard without having become highly proficient at the task.

Using proximate scoring according to the ILR system, we examined the ratings a second time. In this case, the scores ranged from 6-11 with a mean for the group of 9.9. And, although the data base is small, individual patterns could be discerned when proximate and exact scores were compared. Compared to the original ILR ratings, Subject A (Table 2) tended to underrate the difficulty of passages, assigning levels at least a plus point lower on six of the eleven tests.

Subject W demonstrated the opposite tendency, namely overrating three of the eleven passages, underrating one, and rating the remainder exactly as the ILR Testing Committee had done.

Table 2

Assignment of Levels to Reading Passages

Passage	Subject A	B	C	D	E	F	G	H	I	UMinn Mean	ILR Rating
Andres	IM	I	IH	I	A	NH	NH	I	NH	IM	1
Citizens	IM	IH	IH	IH	IL	I	IH	IH	IL	IH	1 +
Frank	5	5	5	5	5	5	5	5	5	5	5
Motor	IM	IH	A	A	IH	I	IH	A	IL	A	2
OK	NH	I	IL	NH	I	NH	NH	N	N	IL	1
Poetry	IL	IH	IM	NH	IH	NH	NH	NH	I	IM	1
Second	A	A	A	A	A	IH	I +	A	IH	A	2
They	A	3	3	3	A +	A	A +	A +	3	3	3
Time	A	A +	A +	A +	A +	A	A	A	A	A +	2 +
TV	NM	N	NM	N	N	NL	NM	N	NM	NM	0 +
When	A +	4	4	4	4	4	4	4 +	4	4	4
= = =	= = =	= = =	= = =	= = =	= = =	= = =	= = =	= = =	= = =	= = =	= = =
Language	Sp	F	Sp	Sp	Sp	G	G	G	F	–	–
ESC Mean	3	8	9	8	6	2	4	5	8	10	11
PSC Mean	8	11	11	11	11	8	11	10	8	11	–

TABLE 2 (Continued)

Assignment of Levels to Reading Passages

	Subject									UMinn Mean	ILR Rating
Passage	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**		
Andres	I	IH	I	–	I	I	NH	I	IH	IM	1
Citizens	IH	IH	A	–	IH	IH	I	IH	IH	IH	1 +
Frank	5	5	5	–	5	5	4 + /5	5	5	5	5
Motor	A	A	IH	–	IH	IH	A	3	A	A	2
OK	I	I	NH	–	NH	NH	NM	I	NH	IL	1
Poetry	I	I	NH	–	I	I	NH	I	I	IM	1
Second	A +	A	IH	–	A +	A +	A	3	A	A	2
They	3	3	3	–	3	3	A +	3	3	3	3
Time	3	3	3 +	–	A +	A +	A +	A +	A +	A +	2 +
TV	NH	NH	N	–	NH	NH	NL	NH	N	NM	0 +
When	4	4	4 +	–	4 +	4 +	A +	4	4	4	4
Language	G	G	Sp	F	F	F	F	Sp	Sp	–	–
ESC Mean	10	9	3	–	7	7	4	9	9	10	11
PSC Mean	11	11	11	–	11	11	9	9	11	11	–

Table 2 (Continued)

Assignment of Levels to Reading Passages

	Subject								
Passage	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**UMinn Mean**	**ILR Rating**
Andres	IM	I	I	IM	I	I	IM	IM	1
Citizens	IH	1 +	IH	A	IM/H	IH	IH	IH	1 +
Frank	5	5	5	5	5	A +	5	5	5
Motor	A	IH	IH	A	A	I	A	A	2
OK	IM	NH/I	I	IM	I	NH	IL	IL	1
Poetry	IM	NM	IH	IM/H	I	IH	IM	IM	1
Second	A	A +	A	A	A +	I	A	A	2
They	4	4 +	3	3 +	4 +	A	3	3	3
Time	4	A	A +	IH	A	A +	A +	A +	2 +
TV	NM	N	N	NM	NH	N	NM	NM	0 +
When	4 +	4	4	4 +	5	A +	4	4	4
Language	G	F	G	F	Sp	F	G	–	–
ESC Mean	7	5	8	5	7	3	10	10	11
PSC Mean	9	10	11	10	9	5	11	11	–

ESC = Exact Scoring
PSC = Proximate Scoring
ESC Mean for all participants: 6.5 Range: 3-10
PSC Mean for all participants: 9.9 Range: 6-11

The information presented in Table 3 and Figure 3 clarifies the direction and extent of deviation. Table 3 displays the mean for ILR ratings on the eleven passages as well as that for each participant, and thus permits comparison. Table 3 also presents the Spearman rho correlation coefficients. To obtain the data in Table 3, the alphabetic designations for subject ratings in Table 2 were assigned numeric equivalents, as shown in Table 4 below.

Table 3

Mean Ratings by Subject Compared to the ILR Mean Ratings with Correlations between Subject and ILR Ratings for the Eleven Passages

SUBJECT	MEAN FOR THE 11 PASSAGES	ILR MEAN	CORRELATION
A	1.918	2.436	.951
B	2.391	2.436	.967
C	2.418	2.436	.984
D	2.300	2.436	.993
E	2.373	2.436	.870
F	1.918	2.436	.991
G	2.073	2.436	.993
H	2.209	2.436	.977
I	1.991	2.436	.926
J	2.527	2.436	.995
K	2.527	2.436	.984
L	2.300	2.436	.965
M	0.000	2.436	.000
N	2.436	2.436	.974
O	2.436	2.436	.974
P	1.945	2.436	.984
Q	2.618	2.436	.956
R	2.391	2.436	.984
S	2.664	2.436	.998
T	2.373	2.436	.954
U	2.391	2.436	.967
V	2.464	2.436	.930
W	2.645	2.436	.977
X	1.755	2.436	.836
Y	2.373	2.436	.993

Figure 3

Plot of Means of Subject Ratings
Showing Deviation from ILR Mean

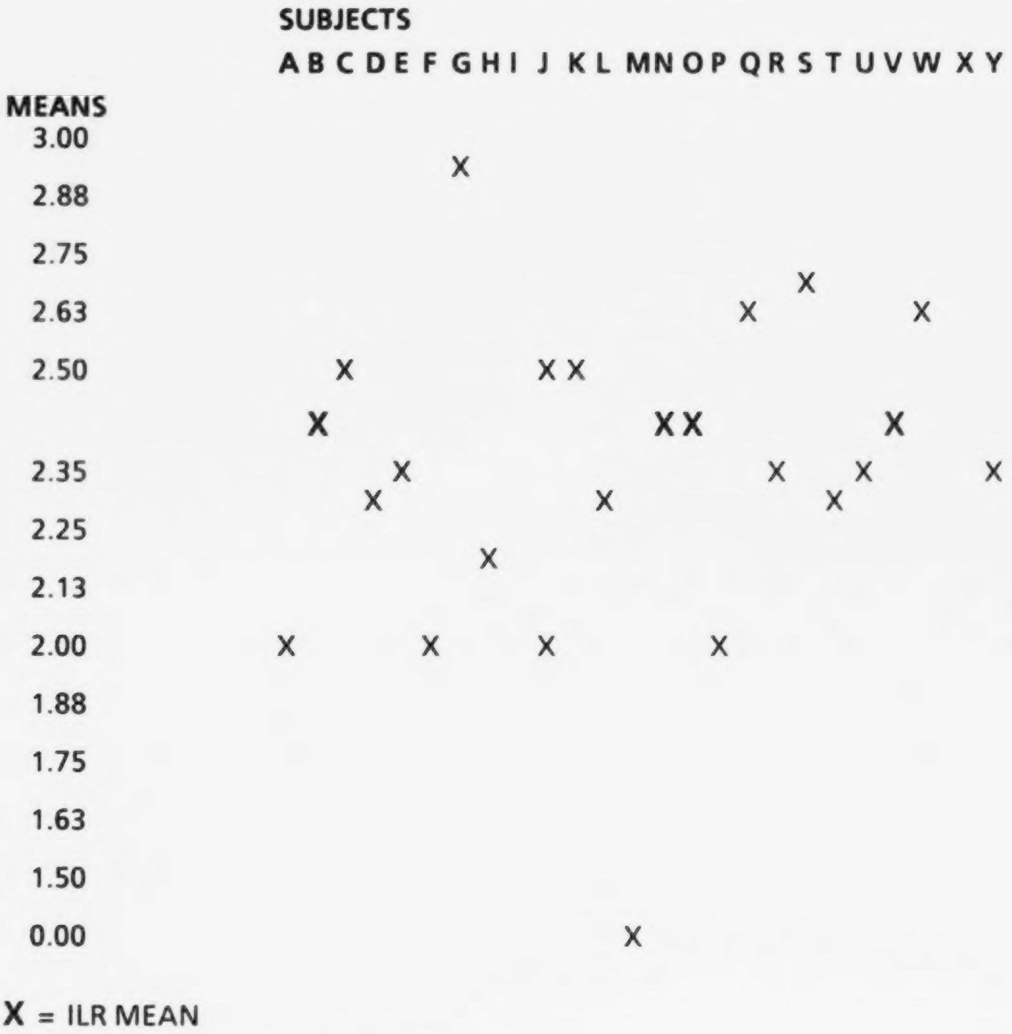


Table 4

Numerical Equivalents
for ACTFL/ETS/ILR Designations

5 = 5.3
4 = 4.3
3 = 3.3
ADV + (2 +) = 2.8
ADV (2) = 2.3
IH (1 +) = 1.8
IM = 1.3
IL = 1.1
I = 1.3
NH = 0.8
NM = 0.3
NL = 0.1
AZ = 0.0

Figure 3 graphs the direction and extent of deviation to which subjects deviated in their ratings from the ILR. The acceptable range spanned 2.25 to 2.57, raising through Proximate Scoring the number of successful raters to 14 in number. Of the 25 participants, 24 completed the rating tasks. Of those Subject X was clearly off the standard. Five participants graded exactly on standard. Subject G graded most leniently, while Subjects Q, S & W were somewhat lenient. Subjects J and K clustered close to the ILR mean. Subjects A, F, H & P were severe. Subjects D, F, L, R, T, U & Y graded rather conservatively, but within the acceptable range.

Thus Table 4 and the Figure 3 permit a closer assessment of each participant's approximation to the ILR standard. Such variation attests both to the difficulty of the task when undertaken individually as well as to the extent of internalization that proved adequate for the group taken as a whole. Again, the goal of the workshop was not to produce fully trained passage graders, but to impart a sense for the system. The figures presented suggest that this goal was achieved.

Conclusions and Concerns

The outcomes of this study are three-fold. First, the ranking task suggests that it is possible to replicate the overall basis for ranking ILR reading passages from easiest to hardest in an academic setting.

Second, the rating task suggests that it is possible to replicate the AEI rating of such passages, matching the passages to suitable AEI definitions. The workshop did not aim at fully training participants in assigning levels to passages, but rather at imparting to them a sense of a functioning system. Fully training individuals would require greater time on the task.

A third conclusion emerges from the participants' workshop evaluations. Participants consistently remarked that the study with its preparatory phase clearly enabled participants to better understand the nature of the ILR assignment of levels to reading passages for proficiency assessment. Discussing the factors contributing to the rating of the twenty-seven ILR sample passages and then checking the extent of the scale's internalization through the tasks of ranking and rating were particularly helpful.

We believe the results of this study are important for any group that is preparing to write tests based on the AEI proficiency statements for reading. Since our findings in this study parallel our experience with internalizing the oral proficiency standard, namely that one must experience proficiency to use the scales accurately, we hazard the conjecture that such training experiences will prove beneficial in proficiency training for all the skill modalities. As a minimum we recommend that such an exercise should begin every item writing workshop for reading proficiency tests.

Even though it appears possible to rate and rank texts with a fair amount of agreement, we express concerns here when choosing texts either for curricular or evaluation purposes. First, individuals bring their own experiences and meaning to a text when they read. The designation of the level of the text should be considered at best an indication of its proficiency level, particularly with more abstract content at more advanced levels. Every text will be of mixed proficiency levels depending on the amount of stylistic variation in the passage and depending on the world knowledge of the readers (Bernhardt, 1986).

Second, the level given to texts in the process described in this study may not necessarily reflect the competence of the individual responding to them. For example, a level may be assigned to a text, but when responding to testing items on the text, examinees may exhibit understanding of the text above, at, or below the passage's level.

We began with the question: Is it possible for potential users to internalize the AEI standards and apply them accurately to grading passages according to the system? The results of this study strongly suggest that such internalization and application are indeed possible.

References

ACTFL Provisional Proficiency Guidelines. 1982. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

ACTFL Proficiency Guidelines. 1986. In Heidi Byrnes and Michael Canale, (eds.), Defining and Developing Proficiency: Guidelines, Implementations, and Concepts. Lincolnwood, IL: National Textbook, pp. 15-24, [Revision of 1982 above.]

Arendt, Jermaine D., Dale L. Lange, and Ray Wakefield. 1986. Strengthening the Language Requirement at the University of Minnesota: An Initial Report. Foreign Language Annals, pp. 149-56.

Bernhardt, Elizabeth B. 1986. Reading in the Foreign Language. In Barbara H. Wing, (ed.), Listening, Reading, Writing: Analysis and Application. Reports of the Northeast Conference on the Teaching of Foreign Languages. Middlebury, VT: The Northeast Conference, pp. 93-115.

Hiple, David V. 1986. A Progress Report on the ACTFL Proficiency Guidelines, 1982-1986. Defining and Developing Proficiency: Guidelines, Implementations, and Concepts. In Heidi Byrnes and Michael Canale, (eds.), Lincolnwood, IL: National Textbook.

Lowe, Pardee Jr. 1986. Proficiency: Panacea, Framework, Process? A Reply to Kramsch, Schulz, and, Particulary, to Bachman and Savignon. Modern Language Journal, 70:391-97.

U.S. Government Interagency Language Roundtable (ILR) Language Skill Level Descriptions: Listening, Speaking, Reading, Writing. Washington, D.C.: Interagency Language Roundtable, 1985. [The descriptions were last revised by Marianne Adams, James Child, John L.D. Clark, Martha Herzog, Pardee Lowe, Jr. with the help of David Hiple, Judith E. Liskin-Gasparro, C. Edward Scebold, and Protase Woodford.]

INTERAGENCY LANGUAGE ROUNDTABLE
LANGUAGE SKILL LEVEL DESCRIPTIONS
READING

Preface

The following proficiency level descriptions characterize comprehension of the written language. Each of the six "base level" (coded 00, 10, 20, 30, 40, and 50) implies control of any previous "plus level" designation (coded 06, 16, 26, etc.) will be assigned when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level." The "plus level" descriptions are therefore supplementary to the "base level" descriptions.

A skill level is assigned to a person through an authorized language examination. Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively.

Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details.

Unless otherwise specified, the term "native reader" refers to native readers of a standard dialect.

"Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education. However, in cultures where formal higher education is common, the language-use abilities of persons who have had such education is considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language.

In the following descriptions a standard set of text-types is associated with each level. The text-type is generally characterized in each descriptive statement.

The word "read," in the context of these proficiency descriptions, means that the person at a given skill level can thoroughly understand the communicative intent in the text-types described. In the usual case the reader could be expected to make a full representation, through summary, or translation of the text into English.

Other useful operations can be performed on written texts that do not require the ability to "read," as defined above. Examples of such tasks which people of a given skill level may reasonably be expected to perform are provided, when appropriate, in the descriptions.

Reading 0 (No Proficiency)

No practical ability to read the language. Consistently misunderstands or cannot comprehend at all. (Has been coded R-0 in some nonautomated applications.) [Data Code 00]

Reading 0 + (Memorized Proficiency)

Can recognize all the letters in the printed version of an alphabetic system and high-frequency elements of a syllabary or a character system. Able to read some or all of the following: numbers, isolated words and phrases, personal and place names, street signs, office and shop designations; the above often interpreted inaccurately. Unable to read connected prose. (Has been coded R-0 + in some nonautomated applications.) [Data Code 06]

Reading 1 + (Elementary Proficiency, Plus)

Sufficient comprehension to understand simple discourse in printed form for informative social purposes. Can read material such as announcements of public events, simple prose containing biographical information or narration of events, and straightforward newspaper headlines. Can guess at unfamiliar vocabulary if highly contextualized, but with difficulty in unfamiliar contexts. Can get some main ideas and locate routine information of professional significance in more complex texts. Can follow essential points of written discussion at an elementary level on topics in his/her special professional field.

In commonly taught languages, the individual may not control the structure well. For example, basic grammatical relations are often misinterpreted, and temporal reference may rely primarily on lexical items as time indicators. Has some difficulty with the cohesive factors in discourse, such as matching pronouns with referents. May have to read materials several times for understanding. (Has been coded R-1 + in some nonautomated applications.) [Data Code 16]

Reading 2 (Limited Working Proficiency)

Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform sorting and locating tasks with written texts that are well beyond their general proficiency level. The individual can read uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straight forward/high-frequency sentence patters. The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. He/she is typically able to answer factual questions about authentic texts of the types described above. (Has been coded R-2 in some nonautomated applications.) [Data Code 20]

Reading 2 + (limited Working Proficiency, Plus)

Sufficient comprehension to understand most factual material in non-technical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to made sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance and/or intentionally disguised meaning. (Has been coded R-2 + in some nonautomated applications.) [Data Code 26]

Reading 3 (General Professional Proficiency)

Able to read within a normal range of speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation, and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas, and "read between the lines," (that is, understand the writers' implicit intents in texts of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms. (Has been coded R-3 in some nonautomated applications.) [Data Code 30]

Reading 3+ (General Professional Proficiency, Plus)

Can comprehend a variety of styles and forms pertinent to professional needs. Rarely misinterprets such texts or rarely experiences difficulty relating ideas or making inferences. Able to comprehend many sociolinguistic and cultural references. However, may miss some nuances and subtleties. Able to comprehend a considerable range of intentionally complex structures, low frequency idioms, and uncommon connotative intentions; however, accuracy is not complete. The individual is typically able to read with facility, understand, and appreciate contemporary expository, technical, or literary texts which do not rely heavily on slang and unusual idioms. (Has been coded R-3+ in some nonautomated applications.) [Data Code 36]

Reading 4 (Advanced Professional Proficiency)

Able to read fluently and accurately all styles and forms of the language pertinent to professional needs. The individual's experience with the written language is extensive enough that he/she is able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references. Able to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment). Able to read and understand the intent of writers' use of nuance and subtlety. The individual can discern relationships among sophisticated written materials in the context of broad experience. Can follow unpredictable turns of thought readily in, for example, editorial, conjectural, and literary texts in any subject matter area directed to the general reader. Can read essentially all materials in his/her special field, including official and professional documents and correspondence. Recognizes all professionally relevant vocabulary known to the educated non-professional native, although may have some difficulty with slang. Can read reasonably legible handwriting without difficulty. Accuracy is often nearly that of a well-educated native reader. (Has been coded R-4 in some nonautomated applications.) [Data Code 40]

Reading 4+ (Advanced Professional Proficiency, Plus)

Nearly native ability to read and understand extremely difficult or abstract prose, a very wide variety of vocabulary, idioms, colloquialisms, and slang. Strong sensitivity to and understanding of sociolinguistic and cultural references. Little difficulty in reading less than fully legible handwriting. Broad ability to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment) is nearly that of a well-read or well-educated native reader. Accuracy is close to that of the well-educated native reader, but not equivalent. (Has been coded R-4+ in some nonautomated applications.) [Data Code 46]

Reading 5 (Functionally Native Proficiency)

Reading proficiency is functionally equivalent to that of the well-educated native reader. Can read extremely difficult and abstract prose; for example, general legal and technical as well as highly colloquial writings. Able to read literary texts, typically including contemporary avantgarde prose, poetry, and theatrical writing. Can read classical/archaic forms of literature with the same degree of facility as the well-educated, but non-specialist native. Reads and understands a wide variety of vocabulary and idioms, colloquialisms, slang, and pertinent cultural references. With varying degrees of difficulty, can read all kind of handwritten documents. Accuracy of comprehension is equivalent to that or a well-educated native reader. (Has been coded R-5 in some nonautomated applications.) [Data Code 50]

A STUDY OF THE COMPARABILITY OF SPEAKING PROFICIENCY INTERVIEW RATINGS ACROSS THREE GOVERNMENT LANGUAGE TRAINING AGENCIES

John L. D. Clark
Defense Language Institute[1]

Background

A pervasive question in the operational use and interpretation of the results of speaking proficiency interviews based on the "ILR" (Interagency Language Roundtable) proficiency level descriptions is the extent to which given examinees' performances would be evaluated in a similar manner across the variety of government agencies and other institutions that make use of this testing procedure. Although there has been a fair amount of conjecture and internal discussion of this topic on the part of examiners and administrators involved in the-day-to-day implementation of agency testing programs, there has not until recently been an opportunity to address the "comparability-of-rating" question in a straightforward empirical manner.

The following is a description of the procedures and major results of a direct experimental comparison of the proficiency ratings assigned to a common group of examinees by testers in each of three government language training agencies: the Language School of the Central Intelligence Agency (CIA), the Defense Language Institute (DLI), and the Foreign Service Institute (FSI). The experiment was conducted using two languages: French and German. Also discussed are the extent to which the results of this particular study might legitimately be extrapolated, cautions on areas in which extrapolation would not be appropriate, and recommendations for follow-up investigation of other aspects of reliability and validity of the interview testing process not formally addressed in the present study.

---

[1] During the period of the study, the author was director of the Foreign Language Education division at the Center for Applied Linguistics (CAL). The assistance of several other CAL staff members in the conduct of the study is gratefully acknowledged. Lynn E. Thompson provided very effective administrative assistance during all phases of the project, Christina Garbacz had major responsibility for data entry as well as for various aspects of statistical processing, and Rebecca Oxford contributed substantially to project planning and procedures specification. In addition to the CAL staff, Nina Levinson (CIA), Thea Bruhn (FSI), and Ellen Mitchell and Phillip White (DLI) coordinated the interviewing activities at their respective agencies with a high level of diligence and effectiveness, and a debt of appreciation is owed the many interviewees in the study who provided, on a voluntary basis, the time and personal interest needed to participate willingly in the interviewing process on three separate occasions. Finally, the major expression of appreciation must be reserved for the certified testers at each of the three agencies, who maintained throughout the six days of testing a seriousness of purpose and diligence of approach to their interviewing and rating tasks that fully demonstrate their high level of professionalism and competence in the proficiency testing role.

Procedure

Overall Study Design

The basic experimental design for the study involved a "test-retest" procedure, in which each examinee was sequentially interviewed by a separate testing team from each of the three participating agencies. In conducting its own interviews, each team made use of the particular interviewing techniques and procedures for arriving at a final rating that were currently in use at that agency. On completion of the process, the team reported a single overall rating on the numerical proficiency scale and associated verbal descriptions of performance endorsed by the ILR member agencies in November 1981 as a "common metric" for speaking proficiency assessment and reporting. This scale comprises six major ratings--0, 1, 2, 3, 4, and 5--supplemented by five intermediate ("plus") ratings--0 +, 1 +, 2 +, 3 +, and 4 +. The scale is intended to characterize the full range of possible learner proficiency levels, from no functional proficiency in the language (Level 0) to proficiency indistinguishable in all respects from that of an educated native speaker (level 5).

Within the administrative and financial constraints involved, it was obviously not possible to carry out such a study for each of the numerous languages in which the agencies routinely test, nor, within a given language, to involve each and every one of the examiners/testers currently conducting interviews within that language. With regard to the selection of languages for the study, discussion with the testing coordinators at each of the three agencies, as well as with the ILR testing subcommittee, resulted in the identification of French and German as two languages for which an adequate number of examinees and testers for the study could be made available within each of the participating agencies, and for which the annual testing volume was sufficiently high to warrant priority attention from an administrative standpoint. With respect to the number of tester teams involved, staff time and travel cost considerations dictated a maximum of two teams per language, for each of the three agencies, i.e., the following configuration:

CIA French Team 1 (all teams are two-person)
CIA French Team 2

DLI French Team 1
DLI French Team 2

FSI French Team 1
FSI French Team 2

CIA German Team 1
CIA German Team 2

DLI German Team 1
DLI German Team 2

FSI German Team 1
FSI German Team 2

Selection of Testers and Examinees

In order to increase the likelihood that, for each agency and language, the testers actually selected for the study would be representative of the total group of individuals operationally testing in that agency/language, the testing coordinator at the agency was asked to provide a complete list of qualified, currently active testers in each language. From this list, the CAL project staff selected all study participants on a statistically random basis. It is thus considered that the composition of the tester groups from each of the three agencies constituted a rigorous random sampling of the population of testers identified by the agency as properly qualified and actively testing within that agency.

A second important design consideration was the selection of examinees. It was considered highly desirable by the project staff, as well as by the ILR testing committee, to investigate rating performance across the full range of proficiency levels by including in the examinee pool individuals covering the gamut from the lowest measurable level (0 +) to the functional equivalent of an educated native speaker (5). At the same time, in view of the fact that the bulk of operational testing at each of the agencies is concentrated within a somewhat smaller band (roughly 1 to 3/3 + for CIA and DLI, 2 to 4 for FSI), it was considered important to insure that a reasonably large number of examinees within this "higher-volume" range would be included in the study sample. To help provide a distribution of examinees for the study that would satisfy both of these criteria, the testing coordinators at each agency were asked to locate and arrange for the participation, per agency, of 20 examinees in each language, and to select these individuals--on the basis of coordinator or language instructor judgments about their proficiency and/or recent interview scores in the agency files--so as to reflect as closely as possible the distribution of proficiency levels shown in Figure 1.

Figure 1

Intended Distribution of Examinee Proficiency Levels at Each Agency

Level	No. of Examinees
0 +	1
1	1
1 +	2
2	3
2 +	3
3	3
3 +	3
4	2
4 +	1
5	1
	20

The coordinators were asked to employ, to the extent possible, a stratified random sampling procedure (for which detailed instructions were given) in identifying the particular examinees who would be asked to participate. The total pool from which the examinees at a given agency were to be drawn was defined to include, in addition to currently-enrolled students, other categories of indiduals that the agency would typically have the occasion to test in the course of its ongoing testing activities (for example, instructor applicants at DLI, or career officers at FSI). Due to a variety of factors, including scheduling conflicts on the part of potential examinees, the necessarily voluntary nature of participation, and the need to locate substitute interviewees on several occasions during the course of the testing, it was not possible to rigorously implement a statistically random process of examinee selection. However, since the major intent in selecting examinees was simply to provide an appropriate overall distribution of proficiency levels across examinees at each agency, some degree of departure from strict random selection of the examinee group was not considered a significant procedural drawback nor an impediment to the proper interpretation of the tester-specific information on which the study was primarily focused.

Scheduling of Interviews

Interviews were conducted on a sequential basis, with two days of testing taking place at each agency. Testing dates were: FSI - September 9-10, 1985; CIA - September 11-12; DLI - September 17-18. On each of these dates, the "home" agency made available all necessary interviewing rooms and other facilities and was responsible for scheduling and contacting the examinees to be tested at that agency by all three tester groups. The examinees at each agency did not themselves travel, but were interviewed on-site by testers from all three agencies on the respective dates above.

Project staff forwarded the testing coordinator at each agency a detailed schedule (See Figure 2) for allocating examiners to testing teams in such a way as to counterbalance the agency-order in which the examinees would be tested as well as to statistically randomize other (uncontrolled) effects attributable to examinees. The test administration schedule, which was followed with only very minimal exceptions at FSI and CIA, involved the administration of all three interviews to a given examinee within a single day. For example, as shown in Figure 2, "examinee 1" was interviewed by a CIA tester team during the first one-hour time period of Day 1, by a DLI team during the third time period, and by an FSI team during the fifth period. Examinee rest breaks of at least one hour were provided between interviews, as well as a one-hour lunch break between either the first and second or second and third interviews. In addition to the lunch period, each tester team had a further one-hour break at some point in the testing day.

Figure 2

Interviewing Schedule

(Cell entries are examinee IDs; same sequence used for French and German)

	Day One						Day Two					
Time Slot	A	B	C	D	E	F	G	H	I	J	K	L
CIA Team 1:	1	2	3	4	5	6	11	12	13	14		
CIA Team 2:	7	8	9	10			15	16	17	18	19	20
DLI Team 1:	5	6	1	2	3	4			11	12	13	14
DLI Team 2:			7	8	9	10	19	20	15	16	17	18
FSI Team 1:	3	4	5	6	1	2	13	14			11	12
FSI Team 2:	9	10			7	8	17	18	19	20	15	16

In setting up the testing schedule described above, it was understood and acknowledged that the per-day "interviewing load" on the part of the testers (six interviews on one day, four on the other) was in some cases more substantial than was typically the case in ongoing testing work at the agency. However, counterbalancing considerations of increased staff costs, additional travel/subsistence expenses, and potential inconvenience on the part of examinees who would be required to appear again on a second or even third testing day, dictated adoption of the former strategy. In a debriefing questionnaire completed at the end of the testing sessions, several examiners reported that they felt somewhat burdened by the overall quantity of interviews required over the available time span, but also for the most part noted that they considered their interviews and associated ratings given in the course of the study to be as thorough and as accurate as those given in regular agency testing.

At DLI, due to restrictions imposed on the scheduling arrangements by both the overall daily schedule at the agency and by individual examinees'classroom session assignments, it was necessary to adopt a somewhat modified procedure in which, for a given examinee, the three interviews were held over a two-day period, on either a 2-1 or 1-2 basis. This modification also permitted a slightly easier and more uniform interviewing pace on the part of the testers, who, with very few exceptions arising from the occasional need to "catch up" for a student who had failed to appear at an assigned testing time, were required to test only 5 students on each of the two days.

Interviewing procedures

All tester teams were extensively advised, both in memoranda circulated prior to the testing and verbally at the beginning of the first testing day, to carry out each interview in strict conformance with the procedures currently in effect at the testers' agency, including, as appropriate, the use of any routine auxiliary materials (e.g., cue cards describing situations that the student is asked to deal with, background reading materials associated with the FSI "briefing" task, and so forth). In addition, the testers were to follow whatever procedures they normally used in arriving at a final interview rating, including, for example, jointly discussing the interviewee's performance; reviewing the verbal proficiency descriptions; and considering (and, if it was the operational procedure at the agency, rating) the speech sample with respect to specified subfactors of performance. Each testing team was also asked to report the final global rating, as well as any factor scores or other routine annotations/feedback information, on the printed forms in use at this agency for this purpose. If separate forms were normally completed by each tester, both were to be submitted; if there was any disagreement concerning the final rating, the testers were to resolve the issue among themselves and circle or otherwise indicate the official "final" rating on one or the other of these forms.

In the course of the interviewing process at all three agencies, the author and another professional project staff member separately sat in on a total of approximately twelve interview sessions, distributed fairly randomly across languages, agencies, and interviewer teams. All interviews conducted by the tester teams, whether or not they were also observed by project staff, were audio recorded on C-90 cassettes, using tape recorders with built-in microphones, with the recorders placed on a table between the examinee and testers. In most instances, the raters' post-interview discussion of the examinee's performance was also recorded. Spot-checking of a number of completed tapes indicated that the spoken material was in general clearly audible with respect to both the examinee and interviewers. The obtained total of over 300 interview recordings is considered to provide a valuable corpus for further linguistic analysis or other follow-up study.

Across all three agencies, 115 examinees were interviewed by testers from each of the three agencies, out of a design total of 120. This very high level of participation is due to both the diligence of the testing coordinators in making the initial administrative arrangments for the interviewing and their willingness and ability to readily locate substitute interviewees as the occasion required over the course of the six testing days.

Results

Overall results

Two types of analysis, chi-square and analysis of variance, were conducted for the testing results as a whole, that is, for the scoring performance of testers across both language groups combined. Table 1 shows the observed and expected frequencies of ratings from 0/0+ (these two levels combined to provide adequate cell size) to 5 on the part of the CIA, DLI, and FSI rating teams. (All tables appear at the end of this chapter.) The overall chi square of 20.3, with a chance probability of .32, fails to demonstrate a statistically significant difference across agencies with respect to the rating of examinee performance on a global (combined languages) basis. It may be noted, however, that the proportion of the total chi square attributable to the FSI rating teams is approximately twice that attributable to either the CIA or DLI testers. As shown in Table 5 below, nonsignificant results ($F = 2.27$; $p = 0.10$) were also obtained for a between- and within-groups analysis of variance (proficiency levels entered as 0+ = 07, 1 = 10, 1+ = 17, etc.) for combined language groups across the three agencies.

Chi-square analyses were also conducted separately for the French (Table 2) and German (Table 3) data. Nonsignificant differences were again found for both languages, with a quite high chance probability for French (.71) and a lower, but still nonsignificant probability (.10) for German. For the French group, the bulk of the (relatively small) chi square is attributable to the ratings given by the FSI testers, but for German, the proportional contribution to chi square is much more evenly distributed across the three agencies. Analysis of variance (Table 5) for the French and German groups considered separately also shows nonsignificant rating differences for both languages across the three participating agencies.

Additional chi-square analyses comparing the rating performance of individual pairs of agencies (CIA-DLI, CIA-FSI, and DLI-FSI) are shown in Table 4 for both whole-group and separate-language comparisons. All of these are statistically nonsignificant ($p > .05$), but it is interesting to note that for both the combined-language and French groups (although not uniformly so for the German group), chance probabilities are appreciably lower for FSI-CIA and FSI-DLI comparisons than for CIA-DLI comparisons. As shown in Table 5, similar results are found for t-test analyses of agency pairs, none of which comparisons reach statistical significance but which show a generally similar patterning of lower chance probabilities for comparisions involving FSI and one of the other two agencies.

In summary of the overall analyses, it may be concluded that the ratings assigned during this study by CIA, DLI, and FSI tester teams, when considered across all the examinee proficiency levels taken as a whole, do not differ among the three agencies or between any pair of agencies in a statistically significant manner, either in combined (French and German) comparisons or in comparisons separately by language.

Inter-Agency Patterns of Score Distribution

Although the whole-group comparisons of scoring performance across the three agencies do not reach statistical significance, examination of the particular scores assigned to examinees within various portions of the overall proficiency range reveals some very interesting patterning. Table 6 shows the interview scores assigned to each examinee by the CIA, DLI, and FSI testers, listed in order of increasing mean score across the three agencies and including both French and German groups. For any given examinee, an asterisk in one of the columns indicates that that particular score is higher than the scores given by both of the other agencies. Of the 115 examinees interviewed, the CIA testers assigned, in 16 instances, a higher score than the other two tester teams. The DLI testers were more generous than their inter-agency colleagues on 8 occasions, and the FSI testers assigned higher scores in 43 cases. A fairly clear pattern is evident in the level 1, 1+, 2 range, with the FSI testers tending in many instances to assign a 1+ (or in a few instances, a 2) to examinees rated as level 1 by CIA and DLI testers. A similar tendency is noted a half-step higher on the scale, with FSI testers assigning level 2 to a number of examinees rated as 1 or 1+ by the other two agencies. A less marked tendency to assign 2 + vis-a-vis 1 + or 2 is also noted.

A tendency on the part of the FSI raters to assign level 3 scores to examinees rated lower than level 3 by the other two agencies is not evident in the data. While there are 6 such instances in the combined French and German data, there are 5 cases in which the CIA testers assigned 3 or 3 + to examinees rated as 2 + or lower by both the DLI and FSI teams. Beyond level 3, the distribution of assigned scores across the three agencies shows generally random differences, with no discernible agency-specific patterning.

Table 7 shows the distribution of ratings across agencies for the French testers separately. The tendency toward relatively higher ratings on the part of the FSI French raters is even more marked than for the combined language group, with higher-than-the-other-two-agency scores assigned by FSI to 30 of the 61 French examinees. Ratings of the CIA French testers were higher than those of their inter-agency colleagues in only 7 instances, and the DLI testers were virtually never higher (2 of 61 occasions) than both of the other agency teams. Rating patterns again show the higher FSI ratings to be most frequent at the 1, 1 +, 2, and 2 + levels. However, in the case of French, there also appears to be some tendency toward the awarding of level 3 scores by FSI to examinees rated as 2 + or lower by the other two agencies (5 instances on the part of the FSI testers, with only one comparable rating by CIA and none by DLI). Inter-agency comparisons of mean interview scores for French show a higher mean value for the FSI ratings (29.5 on a scale defined as level 0 = 00, 0 + = 07, 1 = 10, 1 + = 17, etc.) than for CIA and DLI (means of 25.6 and 25.5, respectively), but this difference is not statistically significant.

Table 8 shows the distribution of German ratings on an across-agencies basis. By contrast to the French data, an apparent tendency to half-point higher ratings on the part of the FSI testers is principally restricted to 1 + vs. 1 and 2 vs. 1 + comparisons, and is by no means as salient or as widespread across proficiency levels as is the case for the French group. Also noteworthy in the German ratings is a tendency to higher ratings on the part of the CIA testers in the middle level of the score range, with level 3 or 3 + assigned by CIA to four examinees rated as 2 + or lower by both DLI and FSI, and 3 + awarded to three other examinees who were considered to be no higher than level 3 by the other two agencies. Mean German interview scores (26.5, 24.9, and 27.5 for CIA, DLI, and FSI, respectively) did not differ significantly across agencies.

Across-agency differences in scoring patterns may also be examined by means of expectancy tables based on the frequencies with which raters from pairs of agencies assigned particular level scores to given examinees. Table 9 shows, for each of the levels assigned by the CIA French testers, the corresponding level assignments of the DLI testers. For example, for the total of 9 interviewees who were rated as level 1 by CIA, 56 percent of these examinees were also rated as level 1 by DLI, 33 percent were rated as level 0 +, and 11 percent, as level 2. For the 10 examinees rated as 1 + by CIA, the DLI ratings were split at 40 percent each for levels 1 + and 2, and 10 percent for levels 1 and 2 +. The discrepancies are more marked for the comparison of CIA and FSI ratings in French (Table 10), which shows, for example, that examinees considered to be at level 1 by the CIA testers were in a majority of cases rated as 1 + (56 percent) by the FSI testers and in a third of the cases, as level 2. At this level, 89 percent of the "level 1" examinees by CIA standards were rated as level

1 + or higher by the FSI testers. This tendency continues through "CIA levels" 1 +, 2, 2 +, 3, and 3 +, with the majority of FSI ratings being at least a half-level higher in all five comparisons. With the exception of "DLI 2 +," comparisons of DLI and FSI French scores (Table 12) reveal an essentially similar pattern across DLI levels 0 through 3, with the bulk of the FSI scores consistently a half-level or more higher than the scores assigned to the same examinees by the DLI testers.

For German, there is no consistent pattern of higher or lower ratings between the CIA and DLI raters from levels 0 + through 2 + (Table 15), but at "CIA levels" 3 and 3 +, the DLI raters were seen to assign somewhat lower ratings on the whole, with an appreciable spread at CIA 3 +, where 50 percent of the corresponding DLI ratings were a full level lower and 20 percent, a level and a half lower. For CIA-FSI comparisions in German (Table 16), there is a clear pattern of at least half-point higher FSI ratings at CIA levels 0 + through 2 +, and a similar pattern for DLI-FSI comparisons (Table 18). A particularly large discrepancy is noted for DLI level 2 +, which shows corresponding FSI scores ranging from 2 to 4 +.

Tables 13-14 (for French) and Tables 19-20 (German) show the variation in scores observed for the CIA and DLI interviews for given score levels on the FSI-conducted interviews. These data may be "read" in the same manner as those shown in the other expectancy tables. For example, as shown in Table 13, of the 12 French interviewees assigned a rating of 2 by the FSI testers, 25 percent received a score of 1 in the interviews conducted by CIA; 50 percent received a rating of 1 +; and 25 percent, a rating of 2.

Three major considerations should be kept in mind in evaluating the observed results. First, at issue in this study is the test-retest reliability of the interviewing process, in which the intent is to determine the extent to which given examinees, undergoing separate, independent interviews by each of the three agencies, will be assigned similar level scores in each instance. Observed variation in examinee score levels may be attributable--in proportions that it is not statistically possible to determine on the basis of the present study--to actual performance differences on the part of the examinee across the three interviewing occasions, as well as to agency-specific differences in the manner in which a given examinee performance would tend to be evaluated across the three agencies. It is, therefore, possible to suggest that at least some of the scoring differences observed in this study may be attributable to interview-to-interview variation in performance on the part of the examinees rather than to rater unreliability per se. However, if the intended operational assumption is that the face-to-face interviewing technique (assuming good will and serious communicative effort on the part of the examinee and diligence and proper attention to elicitation procedures on the part of the examiners) should result in the awarding of similar ratings on closely contemporaneous interviewing occasions, the procedure used in this study may be considered an appropriate empirical approach to determining the validity of this assumption, within the general linguistic and personnel parameters involved (a sampling of two interviewer pairs for two languages across the three participating agencies).

Second, although the total number of interviews obtained in the study was as large as practicable within the financial and administrative constraints involved, and may be considered to provide a reasonably stable and accurate indication of the results that would be secured in a similar but larger study, some caution in interpretation and extrapolation should be exercised, especially in analyzing those expectancy table columns and associated data that are based on a relatively smaller number of interviews.

Third, the expectancy table data should not be viewed as representing in any sense "true" level ratings on the vertical axis. These tables simply show the extent to which the agencies in question tended to vary in the frequencies with which they assigned a given level score to a particular examinee. Any determination of which, if any, of the ratings assigned should be considered to reflect the "true" proficiency level of the examinee is beyond the scope of this study and, indeed, represents a question for which statistical data per se are, at best, of very limited value. Although there is some indication that, for the two languages involved, the interview ratings assigned by CIA and DLI were for certain portions of the overall proficiency scale more similar to each other than they were to the corresponding ratings assigned by FSI, it cannot and should not be concluded from these results that the former ratings were found to be "correct" and the latter "incorrect," in any useful external or criterial sense of the term.

Analysis of Rating "Factor" Data

Some additional information, especially for the interviews conducted by the FSI testing teams, is available concerning the statistical interrelationships of the raters' scoring of various linguistic categories or "factors" that are generally considered to contribute collectively to overall proficiency as expressed in the global rating, but at the same time to provide a certain amount of diagnostic feedback concerning particular sub-aspects of performance (within a given global level) exemplified by a given examinee. Table 21 shows, for combined French and German data, the observed intercorrelations of the FSI global rating and each of the five "factor" scores--"listening comprehension," "discourse," "structure," "lexicalization," and "fluency"-- regularly assigned by FSI testers for interviews conducted by that agency, as an aid in focusing on component aspects of the global rating and in providing for greater objectivity in the rating process overall.

The observed high correlations may be considered attributable to the combined effects of at least two possible sources of correspondence: "true" close relationships among the factors as these are exemplified in the examinee's performance; and a potential "halo effect" arising from the fact that all factor scores for a given examinee are assigned by the same rater, who may be influenced to some extent by examinee performance on one or more of the other factors while attempting to separately and objectively rate a given factor. Although the correlational data suggest that, on a total-group basis, relatively little additional information is provided by the individual factor scores that is not already statistically captured in the global rating, the scoring profiles of particular examinees whose pattern of factor ratings shows an appreciable departure from linearity may be of interest from a diagnostic or pedagogical standpoint. For example, the scatterplot of "structure" vs. "lexicalization" scores shown in Table 22 shows three examinees whose factor ratings for "structure" were proportionately appreciably higher than their ratings for "lexicalization"; and three other examinees for whom the "lexicalization" scores were noticeably higher than the "structure" scores. Although detailed linguistic review of the interview performance of particular examinees is beyond the scope of the present report, the scoring data obtained in the study can serve to identify these and other "discrepant" cases for further clinical analyses addressing, for example, the so-called "street learner"/"school learner" performance differences frequently reported in operational testing activities.

Detailed factor score data are not available for the CIA or DLI interviews in that, for the most part, testers from these two agencies followed the current operational procedure of providing only the overall global rating, with the single exception of a separate "listening comprehension" score that was consistently awarded by by the CIA raters and in about two-thirds of the cases by the DLI raters. The obtained "listening" vs. "global" correlations were .97 for the CIA interviews (N = 114) and .98 for the DLI interviews (N = 85). These data again suggest that, on a whole-group basis, very little "new" information is provided by the separate listening score. Analysis of individual discrepant cases for clinical or pedagogical purposes would of course be possible for the CIA and DLI data as well as for the FSI interviews.

Examinee and Tester Feedback on Interviewing Process

The observations and opinions of both examinees and interviewers concerning various aspects of the interviewing procedures as exemplified during the study were solicited through two separate questionnaires (Appendices A and B). The examinee questionnaire requested information on the examinee's affiliation and test language, and both "yes-no" and open-ended responses to the following questions for each of the three interviews taken:

"Did the opportunities the testers provided you to speak the language during the [first, second, third] interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?"

"Did the testers during the [first, second, third] interview use any elicitation techniques or cover any kinds of topics that you thought were in any way "unfair" or in some other way not a valid test of your speaking proficiency?"

"During the [first, second third] interview, did the testers appear to make a conscious effort to put you at ease?"

Four additional summary questions involved forced-choice judgments as follows:

"In which of the three interviews do you feel you . . .

were most relaxed and at ease?

were the most anxious or nervous?

best demonstrated your optimum speaking proficiency?

least well demonstrated your optimum speaking proficiency?"

Questionnaires were distributed to the examinees by the testing coordinator at each agency within about one week following completion of the interviewing process, an approach intended to avoid the possibility that examinees filling out the questionnaire immediately on completion of the testing might be disproportionately influenced by their experience in the most recently-taken interview. The examinee was asked to provide his or her name on an attached slip to permit the project staff to properly categorize the 'first,' 'second,' and 'third' interviews taken, but was assured that the slip would be removed when the results were summarized and that all data would be analyzed and reported on an anonymous basis. A self-addressed, postpaid envelope was provided for return of the questionnaire. Of the 115 examinees participating in the study, questionnaires were returned by 83, a response rate of 72 percent.

Table 23 provides a summary of the examinee questionnaire responses. To the question "Did the opportunities the testers provided you to speak the language during the interview...adequately probe your maximum proficiency level?," by far the greatest number of responses (87 percent overall) were in the affirmative. Chi-square analysis for CIA, DLI, and FSI interviews showed no significant differences across agencies in the frequency with which the examinees reported an adequate probing of maximum proficiency level. To the question of elicitation techniques or coverage of topics that the examinee considered "'unfair' or in some other way not a valid test of your speaking proficiency," 82 percent of the total judgments across agencies were that "unfair" techniques had not been used. However, on an agency-specific basis, the corresponding chi-square is highly significant (p = .007) with 29 percent of the FSI interviews being judged as "unfair" in procedure or topical coverage, as contrasted to 13 percent and 12 percent for the CIA and DLI interviews, respectively.

The question "Did the testers appear to make a conscious effort to put you at ease?" was answered affirmatively in almost 9 out of 10 cases overall (89 percent), but chi-square analysis again shows a significant across-agency difference (p = .024), with a somewhat smaller proportion of the FSI interviews (82 percent) being judged as consciously directed toward putting the examinee at ease, by comparison to the corresponding CIA (89 percent) and DLI (95 percent) figures.

Although the total number of data elements for the forced-choice questions (one rather than three per examinee) are insufficient for across-agency statistical comparison, the absolute frequencies of response to these questions appear to corroborate rather closely the results of the earlier questions. To the question, "In which of the three interviews do you feel you were most relaxed and at ease?," 40 interviewees indicated "DLI"; 27, "CIA"; and 12, "FSI". The conversely-phrased question, "In which of the three interviews do you feel you were the most anxious or nervous?," showed even greater differentiation across agencies, with only 9 interviewees identifying "DLI"; 15, "CIA"; and 50, "FSI." Notwithstanding an apparent clear discrimination on the examinees' part as to the relative ease/anxiety producing qualities of the interview as conducted by each of the three agencies, no appreciable across-agency differences are shown in their judgments of the agency providing the best or worst opportunity to demonstrate their optimum speaking proficiency.

To determine possible differences in questionnaire response tendencies attributable to an interaction between the agency affiliation of the examinees and that of the tester teams--that is, to investigate the possibility that, for example, DLI students might have reported different experiences or opinions concerning their participation in the DLI-conducted interviews than did interviewees from CIA or FSI being tested by the DLI teams (or analogously for other examinee/agency combinations)--additional chi square analyses of each of the questions summarized in Table 23 were carried out for the crosstabulations of interviewee agency and tester agency. All of these analyses showed nonsignificant ($p > .05$) interaction effects, suggesting that reported examinee reactions to their experiences in being tested by each of the three agencies did not vary to any meaningful extent as a consequence of their own agency affiliation. In interpreting these results, it is necessary to note that at all three agencies, a small number of examinees were necessarily drawn from non-student sources (e.g., agency alumni) in order to provide an adequate sample at the top end of the proficiency scale. Reactions of these individuals to the interviewing process may have been somewhat atypical of those of current students; however, to have excluded these non-student cases from the interaction analysis would have reduced the already small cell sizes to statistically insufficient levels.

The questionnaire completed by the examiners themselves (Appendix B) was somewhat less formal than the examinee questionnaire and requested open-ended comments by the testers concerning several aspects of their interviewing in the course of the project. Of the 24 testers participating in the study, 18 returned completed questionnaires (75 percent). Responses were on an intentionally anonymous basis, with only the tester's "language and agency affiliation" being requested on the questionnaire form. As shown in Table 24, based on the project staff's categorizations for analysis purposes of the free-response answers, the great majority of testers felt that the interviewing procedures they had used during the study were the same as those used during "routine, day-to-day testing" at their agency; and that the ratings which they assigned were, on the whole, as accurate as those typically made during operational testing. To the question, "Do you feel that the accuracy of your ratings varied at certain times or points during the six-day testing period?," most respondents were of the opinion that their judging accuracy had not varied appreciably over the course of the testing, but some cited the relatively intensive testing schedule (involving in some cases up to six interviews per day) as a potential source of end-of-day fatigue and consequent lack of full and "fresh" attention to the interviewing and rating tasks. With respect to the possible effects of "examiner fatigue" on the overall study results, it should be emphasized that the counterbalanced scheduling of the interviewing sessions was intended to adjust operationally for this and other possible sequence-of-interviews-related factors insofar as the inter-agency comparisons at issue in the study are concerned.

Some differences in the overall composition of the examinee groups at the three different agencies were also noted by the testers, with the FSI and CIA examinees, in general, considered to be more highly qualified on a total-group basis than the DLI interviewees. Again, the balanced nature of the study design, in which testers from each agency interviewed the same examinees at all three testing locations, would be expected to rule out any effects of inter-agency differences in examinee populations with respect to the project results per se.

Despite the fairly rigorous testing schedule, which involved both concentrated interviewing on a day-to-day basis and travel between Washington and Monterey within a relatively brief time span, the great majority of interviewers felt that their participation in the project had been of interest and benefit to them. Cited especially in this regard were the opportunities to meet and interact with testers from other agencies and to "share notes" on both personal and professional basis. Several examiners expressed the hope that similar projects undertaken in the future could have built into them more extensive and more formally-structured opportunities for this type of interaction.

Summary

The major result of the study may be summarized as follows. With respect to the testing of French and German by trained CIA, DLI, and FSI interviewer/raters, as represented by two randomly selected two-person teams for each agency and language, who interviewed and rated a total of 20 examinees each across essentially the full spectrum of proficiency levels, the ratings assigned did not differ across agencies in a statistically significant way, either on a combined (French plus German) or individual-language basis. Notwithstanding these overall results, examination of the rating performance for various sub-portions of the proficiency scale showed fairly clear across-agency differences for both languages, primarily at the lower and middle ranges of the scale, with these differences for the most part reflecting relatively higher rating assignments on the part of the FSI raters by comparison to the ratings given by the other two agencies. As shown both in the distribution of test scores for the same examinees across agencies and in a series of two-way expectancy tables derived from these distributions, there are occasional fairly wide discrepancies in scoring for individual examinees, which suggests the advisability, on a follow-up basis, of clinically studying the most discrepant cases from both linguistic and interviewing-procedure standpoints, to attempt to identify common factors that may have contributed to these scoring differences.

Analysis of the intercorrelations of the FSI "factor" scores among themselves and with the global ratings shows very high correspondence among all of these variables. Correlations of the CIA and FSI "listening" scores with the global ratings were also extremely high. These results suggest that, notwithstanding the possible utility of the factor scoring process in facilitating the interviewers' overall rating task, relatively little new or different statistical information is provided by the factor scores by comparison to the information already contained in the global ratings. However, factor score analysis does make it possible to identify individual examinees showing atypical (non-linear) factor score patterns, and detailed linguistic analysis of the interview performance of individuals showing such patterns may be of both pedagogical and research interest.

Questionnaire-based information obtained from the participating examinees indicates that, for the most part, the examinees felt that their optimum level of proficiency had been adequately probed in interviews conducted by all three agencies. There were, however, appreciable differences in the examinees' affective reactions to the interviewing process, with a statistically significant tendency for the the examinees to view the FSI interviewing procedure as both more anxiety-producing and making use of what they considered to be "unfair" elicitation techniques. Also on the basis of questionnaire responses, the great majority of participating testers reported that in their opinion, the interviews which they had conducted during the study were quite similar to the operational interviews given at their home agency with respect to interviewing procedures and accuracy of ratings, although the atypically long testing day was cited in some instances as a potential source of differences in both areas. Virtually all testers found their own involvement in the study to have been quite rewarding from personal and/or professional standpoints.

With regard to extrapolation of study results, it is reasonable to assume, as a consequence of the experimental design, that the group of testers selected to participate in the study constituted a random sample of the population of testers currently interviewing at each agency in the languages in question. As such, their performance may be considered indicative of the probable total group characteristics of testers in that language/agency combination, without, however, ruling out the possibility that the "luck of the draw" may have in some instances placed in the sample individuals having atypical characteristics in terms of their elicitation procedures or accuracy of rating vis-a-vis those of their colleagues.

Considerable caution should be exercised in extrapolating the observed results for French and German testing to other languages not formally investigated in the study, both in view of the fact that the non-studied languages have different populations of testers, and in consideration of possible linguistically-based differences across languages that would have an operational bearing on the interviewing process and/or on the reliability of the ratings assigned. It should also be emphasized that the present study provides information about the test-retest comparability of the interviewing process on an across-agencies basis, and does not directly examine the question of rating reliability within a given agency (i.e., the extent to which each of several raters within one agency would agree with one another in repetitive interviewing of a given examinee), and it is quite possible to suggest that the level of scoring agreement within any one agency would be greater than that observed on an inter-agency basis. However, to the extent that the ILR scale- based interview is intended to represent a "common metric" of examinee performance, with identical meaning and interpretation across using agencies, the results of the present study warrant close examination for possible conceptual or procedural implications that would arise from holding such an objective.

Table 1

Chi-Square by Agency and Interview Score Assigned

(French and German)

Observed Expected (O - E) Contribution	CIA	DLI	FSI	Row Totals
0, 0 +	8 6.7 1.3 0.3	9 6.7 2.3 0.8	3 6.7 -3.7 2.0	20 3.1
1	19 14.7 4.3 1.3	15 14.7 0.3 0.0	10 14.7 -4.7 1.5	44 2.8
1 +	16 14.0 2.0 0.3	17 14.0 3.0 0.6	9 14.0 -5.0 1.8	42 2.7
2	12 15.3 -3.3 0.7	13 15.3 -2.3 0.4	21 15.3 5.7 2.1	46 3.2
2 +	10 14.7 -4.7 1.5	19 14.7 4.3 1.3	15 14.7 0.3 0.0	44 2.8
3	10 13.0 -3.0 0.7	11 13.0 -2.0 0.3	18 13.0 5.0 1.9	39 2.9
3 +	16 12.7 3.3 0.9	11 12.7 -1.7 0.2	11 12.7 -1.7 0.2	38 1.3
4	10 10.7 -0.7 0.0	9 10.7 -1.7 0.3	13 10.7 2.3 0.5	32 0.8

Table 1 (Continued)

4 +	8 7.7 0.3 0.0	6 7.7 -1.7 0.4	9 7.7 1.3 0.2	23 0.6
5	6 5.7 0.3 0.0	5 5.7 -0.7 0.1	6 5.7 0.3 0.0	17 0.1
Column Totals	115 5.7	115 4.3	115 10.3	345 20.3

No. of Observations = 345 Degrees of freedom = 18

Chi square = 20.3 Chance probability = 0.32

Table 2

Chi-Square for French Interviews

Observed Expected (O - E) Contribution	**CIA**	**DLI**	**FSI**	**Row Totals**
0, 0 + , 1	13 10.3 2.7 0.7	13 10.3 2.7 0.7	5 10.3 -5.3 2.8	31 4.1
1 +	10 8.0 2.0 0.5	9 8.0 1.0 0.1	5 8.0 -3.0 1.1	24 1.8
2	8 9.0 -1.0 0.1	7 9.0 -2.0 0.4	12 9.0 3.0 1.0	27 1.6
2 +	7 7.0 0.0 0.0	7 7.0 0.0 0.0	7 7.0 0.0 0.0	21 0.0
3	4 6.3 -2.3 0.9	7 6.3 0.7 0.1	8 6.3 1.7 0.4	19 1.4
3 +	6 6.3 -0.3 0.7	5 6.3 -1.3 0.3	8 6.3 1.7 0.4	19 0.7
4	5 6.7 -1.7 0.4	7 6.7 0.3 0.0	8 6.7 1.3 0.3	20 0.7
4 + , 5	8 7.3 0.7 0.1	6 7.3 -1.3 0.2	8 7.3 0.7 0.1	22 0.4
Column Totals	61 2.7	61 1.9	61 6.1	183 10.6

No. of Observations = 183
Chi square = 10.6

Degrees of freedom = 18
Chance probability = 0.71

Table 3

Chi-Square for German Interviews

Observed Expected (O - E) Contribution	CIA	DLI	FSI	Row Totals
0, 0 +, 1	14	11	8	
	11.0	11.0	11.0	33
	3.0	0.0	-3.0	1.6
	0.8	0.0	0.8	
1 +	6	8	4	
	6.0	6.0	6.0	18
	0.0	2.0	-2.0	1.3
	0.0	0.7	0.7	
2	4	6	9	
	6.3	6.3	6.3	19
	-2.3	-0.3	2.7	2.0
	0.9	0.0	1.1	
2 +	3	12	8	
	7.7	7.7	7.7	23
	-4.7	4.3	0.3	5.3
	2.8	2.4	0.0	
3	6	4	10	
	6.7	6.7	6.7	20
	-0.7	-2.7	3.3	2.8
	0.1	1.1	1.7	
3 +	10	6	3	
	6.3	6.3	6.3	19
	3.7	-0.3	-3.3	3.9
	2.1	0.0	1.8	
4, 4 +, 5	11	7	12	
	10.0	10.0	10.0	30
	1.0	-3.0	2.0	1.4
	0.1	0.9	0.4	
Column Totals	54	54	54	162
	6.8	5.1	6.4	18.4

No. of Observations	=	162	Degrees of freedom	=	12
Chi square	=	18.4	Chance probability	=	0.10

Table 4

Chi-Square for Agency Pairs

		X2	N	df	p
French and German					
	CIA - DLI	4.8	230	9	0.85
	CIA - FSI	14.1	230	9	0.12
	DLI - FSI	11.9	230	9	0.22
French					
	CIA - DLI	1.6	122	7	0.98
	CIA - FSI	8.3	122	7	0.30
	DLI - FSI	7.1	122	7	0.42
German					
	CIA - DLI	8.7	108	6	0.19
	CIA - FSI	11.0	108	6	0.09
	DLI - FSI	8.1	108	6	0.23

Table 5

Analysis of Variance and t-Test Comparisons for Interview Scores across Three Agencies

Source of Variation	df	Sum of Squares	Mean Square	F	P
French and German					
Between groups	2	722.649	361.325	2.27	0.10
Within groups	342	54477.078	159.290		
Total	344	55199.728			
French					
Between groups	2	648.995	324.497	2.00	0.14
Within groups	180	29232.361	162.402		
Total	182	29881.355			
German					
Between groups	2	192.704	96.352	0.61	0.55
Within groups	159	25098.241	157.851		
Total	161	25290.944			

Table 5 (Continued)

t-Statistics

	Comparison	t	p
French and German			
	CIA - DLI	.517	0.64
	CIA - FSI	1.531	0.22
	DLI - FSI	2.048	.13
French			
	CIA- DLI	.078	0.94
	CIA - FSI	1.691	0.19
	DLI - FSI	1.769	0.18
German			
	CIA - DLI	.674	0.55
	CIA - FSI	.421	0.70
	DLI - FSI	1.095	0.36

Table 6

Examinee Score Levels Assigned, by Agency
(French and German)

Legend: 00 = 0, 07 = 0 +, 10 = 1, 17 = 1 +, etc.
Asterisks indicate a score higher than that of the other two agencies.

CIA	DLI	FSI
07	00	07
07	07	07
07	07	10*
07	10*	07
07	07	10*
07	07	10*
07	10	10
10	10	10
10	10	10
10	10	10
10	10	10
17*	07	10
10	07	17*
10	07	17*
07	17*	10
10	07	17*
10	10	17*
10	10	17*
10	10	17*
10	10	17*
10	10	17*
10	10	20*
10	10	20*
10	17	17
10	17	20*
17	10	20*
10	17	20*
17	10	20*

Table 6 (Continued)

10	20	20
10	17	27*
17	17	20*
17	17	20*
17	17	20*
17	17	20*
17	17	20*
20	17	20
17	20	20
17	20	20
20	17	20
20	20	20
17	17	27*
17	20	27
20	17	27*
17	20	27*
17	27*	20
20	17	27*
17	20	30*
20	27*	20
17	20	30*
20	20	30*
20	27	27
27	17	30*
20	27	27
20	27	27
20	27	27
30*	20	27
30*	20	27
20	30	30
27	27	27
27	27	30*
30*	27	27
37*	17	30
27	27	30*
37*	27	20

Table 6 (Continued)

37*	27	20
27	30	30
27	30	30
30	27	30
30	30	30
37*	27	30
30	27	37*
37*	27	30
27	37*	30
37*	27	30
37	20	37
27	30	37*
37	20	37
30	30	37*
40*	30	27
27	30	40*
30	30	40*
27	37	37
37	37	30
37	37	30
37	30	40*
40*	30	37
30	37	40*
40	27	40
37	37	37
37	37	40*
40*	37	37
40*	37	37
40	40	37
37	40	40
37	40	40
30	47*	40
40	40	40
47	27	47
37	40	47*
40	47*	40

Table 6 (Continued)

	47*	40	40
	40	40	47*
	47*	40	40
	47	37	47
	47	37	47
	40	47	47
	47	40	47
	47	50*	47
	47	50	50
	50	47	50
	50	50	47
	50	47	50
	50	47	50
	50	50	50
	50	50	50
Mean:	26.0	25.2	28.6
S.D.:	13.4	12.5	11.8
N:	115	115	115

Table 7

Examinee Score Levels Assigned, by Agency
(French)

Legend: 00 = 0, 07 = 0 +, 10 = 1, 17 = 1 +, etc.
Asterisks indicate a score higher than that of the other two agencies.

CIA	DLI	FSI
07	00	07
07	07	10*
07	07	10*
07	07	10*
10	10	10
10	07	17*
10	07	17*
10	07	17*
10	10	17*
10	10	17*
10	10	20*
10	10	20*
17	10	20*
10	20	20
17	17	20*
17	17	20*
17	17	20*
20	17	20
17	20	20
20	17	20
20	20	20
17	17	27*
20	17	27*
17	20	27*
17	27*	20
17	20	30*
17	20	30*
20	20	30*
20	27	27

Table 7 (Continued)

	27	17	30*
	20	27	27
	20	30	30
	27	27	27
	27	27	30*
	30*	27	27
	37*	17	30
	27	30	30
	30	27	37*
	27	30	37*
	30	30	37*
	27	30	40*
	27	37	37
	37	30	40*
	40*	30	37
	37	37	37
	40*	37	37
	40*	37	37
	37	40	40
	37	40	40
	30	47*	40
	37	40	47*
	40	47*	40
	47*	40	40
	40	40	47*
	47*	40	40
	47	37	47
	47	40	47
	47	50	50
	50	50	47
	50	50	50
	50	50	50
Mean:	25.6	25.5	29.5
S.D.:	13.2	13.3	11.5
N:	61	61	61

Table 8

Examinee Score Levels Assigned, by Agency (German)

Legend: 00 = 0, 07 = 0 +, 10 = 1, 17 = 1 +, etc.
Asterisks indicate a score higher than that of the other two agencies.

CIA	DLI	FSI
07	07	07
07	10	07
07	10	10
10	10	10
10	10	10
10	10	10
17*	07	10
07	17*	10
10	10	17*
10	10	17*
10	10	17*
10	17	17
10	17	20*
17	10	20*
10	17	20*
10	17	27*
17	17	20*
17	17	20*
17	20	20
17	20	27*
20	17	27*
20	27*	20
20	27	27
20	27	27
30*	20	27
30*	20	27
27	27	30*
37*	27	20

Table 8 (Continued)

	37*	27	20
	27	30	30
	30	27	30
	30	30	30
	37*	27	30
	37*	27	30
	27	37*	30
	37*	27	30
	37	20	37
	37	20	37
	40*	30	27
	30	30	40*
	37	37	30
	37	37	30
	30	37	40*
	40	27	40
	37	37	40*
	40	40	37
	40	40	40
	47	27	47
	47	37	47
	40	47	47
	47	50	47
	50	47	50
	50	47	50
	50	47	50
Mean:	26.5	24.9	27.5
S.D.:	13.6	11.7	12.0
N:	54	54	54

Table 9

Expectancy Table for DLI from CIA Scores

(French)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by CIA.

		DLI											
		0	0 +	1	1 +	2	2 +	3	3 +	4	4 +	5	N
	0 +	25	75										4
	1		33	56		11							9
	1 +			10	40	40	10						10
C	2				38	25	25	13					8
I	2 +				14		29	43	14				7
A	3						50	25			25		4
	3 +				17			17	17	50			6
	4							20	40	20	20		5
	4 +								20	60		20	5
	5											100	3

Table 10

Expectancy Table for FSI from CIA Scores

(French)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by CIA.

		FSI										
		0 +	1	1 +	2	2 +	3	3 +	4	4 +	5	N
	0 +	25	75									4
	1		11	56	33							9
	1 +				60	20	20					10
C	2				38	38	25					8
I	2 +					14	43	29	14			7
A	3					25		50	25			4
	3 +						17	17	50	17		6
	4							60	20	20		5
	4 +								40	40	20	5
	5									33	67	3

Table 11

Expectancy Table for CIA from DLI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by DLI.

		CIA										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0	100										1
	0+	50	50									6
	1		83	17								6
D	1+			44	33	11		11				9
L	2		14	57	29							7
I	2+			14	29	29	29					7
	3				14	43	14	14	14			7
	3+					20		20	40	20		5
	4							43	14	43		7
	4+						50		50			2
	5									25	75	4

Table 12

Expectancy Table for FSI from DLI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by CIA.

		FSI										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0	100										1
	0+		50	50								6
	1		17	33	50							6
	1+				56	22	22					9
D	2				43	14	43					7
L	2+				14	57	14	14				7
I	3						29	43	29			7
	3+							80		20		5
	4								57	43		7
	4+								100			2
	5									25	75	4

Table 13

Expectancy Table for CIA from FSI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by FSI.

		CIA										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	100										1
	1	75	25									4
	1+		100									5
F	2		25	50	25							12
S	2+			29	43	14	14					7
I	3			25	25	38		13				8
	3+					25	25	13	38			8
	4					13	13	38	13	25		8
	4+							20	20	40	20	5
	5									33	67	3

Table 14

Expectancy Table for DLI from FSI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by FSI.

		DLI											
		0	0 +	1	1 +	2	2 +	3	3 +	4	4 +	5	N
	0 +	100											1
	1		75	25									4
	1 +		60	40									5
F	2			25	42	25	8						12
S	2 +				29	14	57						7
I	3				25	38	13	25					8
	3 +						13	38	50				8
	4							25		50	25		8
	4 +								20	60		20	5
	5											100	3

Table 15

Expectancy Table for DLI from CIA Scores

(German)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by CIA.

		DLI										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	25	50	25								4
	1		60	40								10
	1+	17	17	33	33							6
C	2			25		75						4
I	2+					33	33	33				3
A	3				33	17	33	17				6
	3+				20	50		30				10
	4					20	20		40	20		5
	4+					33		33			33	3
	5									100		3

Table 16

Expectancy Table for FSI from CIA Scores

(German)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by CIA.

		FSI										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	50	50									4
	1		30	40	20	10						10
	1+		17		67	17						6
C	2				25	75						4
I	2+						100					3
A	3					33	33		33			6
	3+				20		50	20	10			10
	4					20		20	40	20		5
	4+									100		3
	5										100	3

Table 17

Expectancy Table for CIA from DLI. Scores

(German)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by DLI.

CIA

		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	50		50								2
	1	22	67	11								9
	1+	13	50	25	13							8
D	2			33			33	33				6
L	2+				25	8	8	42	8	8		12
I	3					25	50		25			4
	3+					17	17	50		17		6
	4								100			2
	4+								25		75	4
	5									100		1

Table 18

Expectancy Table for FSI from DLI. Scores

(German)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by DLI.

		FSI										
		0 +	1	1 +	2	2 +	3	3 +	4	4 +	5	N
	0 +	50	50									2
	1	11	44	33	11							9
	1 +		13	13	50	25						8
D	2				17	50		33				6
L	2 +				25	17	42		8	8		12
I	3					25	50		25			4
	3 +						50		33	17		6
	4							50	50			2
	4 +									25	75	4
	5									100		1

Table 19

Expectancy Table for CIA from FSI Scores

(German)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by FSI.

		CIA										
		0 +	1	1 +	2	2 +	3	3 +	4	4 +	5	N
	0 +	100										2
	1	33	50	17								6
	1 +		100									4
F	2		22	44	11			22				9
S	2 +		13	13	38		25		13			8
I	3					30	20	50				10
	3 +							67	33			3
	4						40	20	40			5
	4 +								25	75		4
	5										100	3

Table 20

Expectancy Table for DLI from FSI Scores

(German)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by FSI.

		DLI										
		0 +	1	1 +	2	2 +	3	3 +	4	4 +	5	N
	0 +	50	50									2
	1	17	67	17								6
	1 +		75	25								4
F	2		11	44	11	33						9
S	2 +			25	38	25	13					8
I	3					50	20	30				10
	3 +				67				33			3
	4					20	20	40	20			5
	4 +					25		25		25	25	4
	5									100		3

Table 21

Intercorrelations of FSI Factor Ratings

(N = 115)

	Global Rating	Comprehension	Discourse	Structure	Lexicalization
Comprehension	.94				
Discourse	.97	.92			
Structure	.97	.92	.96		
Lexicalization	.97	.94	.95	.96	
Fluency	.95	.94	.94	.93	.94

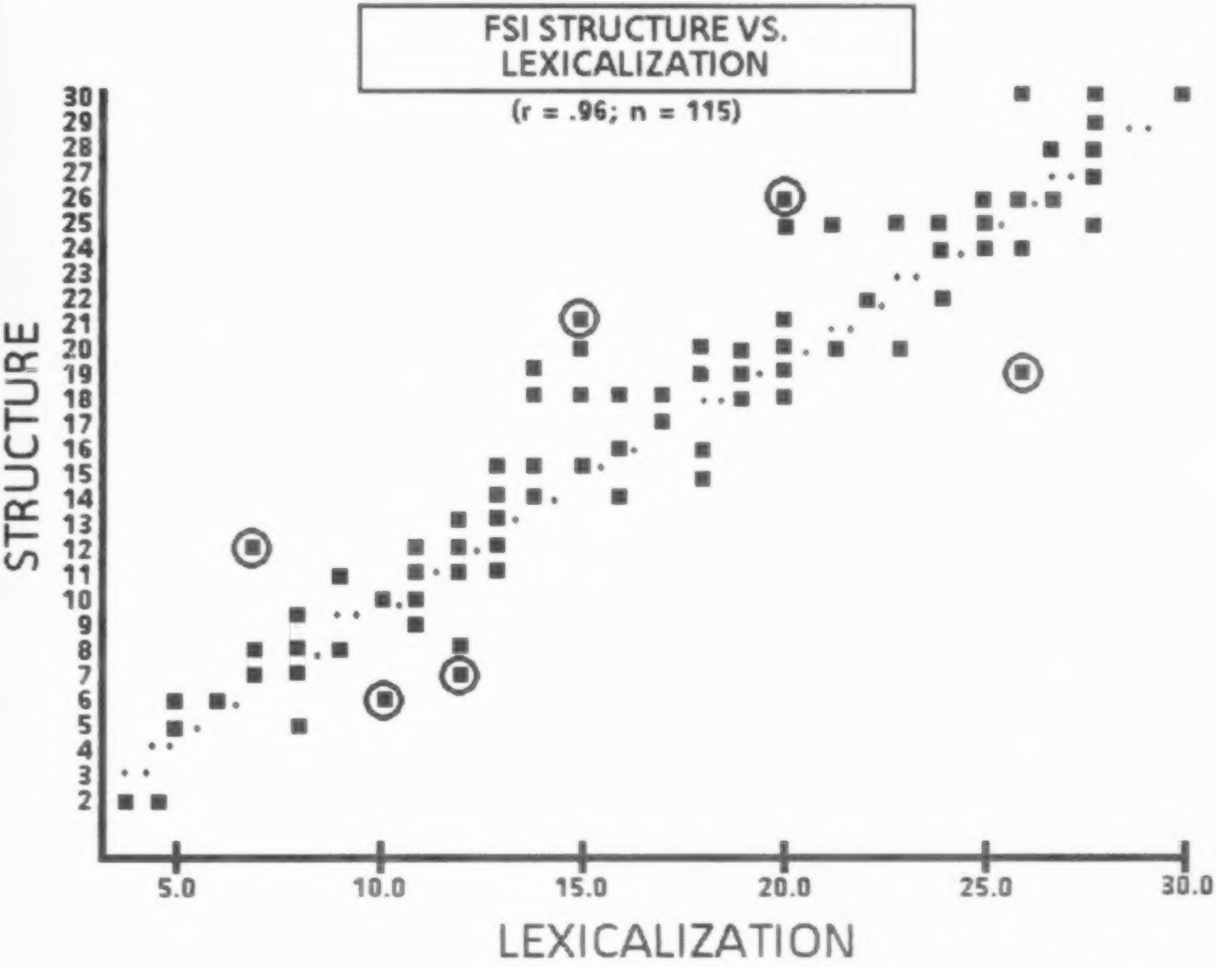

FSI STRUCTURE VS. LEXICALIZATION
(r = .96; n = 115)
STRUCTURE
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
5.0
10.0
15.0
20.0
25.0
30.0
LEXICALIZATION

Table 23

Summary of Responses to Examinee Questionnaire

"Did the opportunities the testers provided you to speak the language during the interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?"

	CIA Interview	DLI Interview	FSI Interview	Total
YES	90%	86%	85%	87%
NO	10%	14%	15%	13%
Total Responses:	72	72	75	219

Chi square = .92; p = .63

"Did the testers use any elicitation techniques or cover any kinds of topics that you thought were in any way "unfair" or in some other way not a valid test of your speaking proficiency?"

	CIA	DLI	FSI	Total
YES	13%	12%	29%	18%
NO	87%	88%	71%	82%
Total Responses:	83	83	83	249

Chi square = 9.93; p = .007

Table 23 (Continued)

"Did the testers appear to make a conscious effort to put you at ease?"

	CIA	DLI	FSI	Total
YES	89%	95%	82%	89%
NO	11%	5%	18%	11%
Total Responses:	83	83	82	248

Chi square = 7.50; p = .024

"In which of the three interviews do you feel you were most relaxed and at ease?"

CIA	DLI	FSI
27	40	12

"In which of the three interviews do you feel you were the most anxious or nervous?"

CIA	DLI	FSI
15	9	50

"In which of the three interviews do you feel you best demonstrated your optimum speaking proficiency?"

CIA	DLI	FSI
23	29	25

"In which of the three interviews do you feel you least well demonstrated your optimum speaking proficiency?"

CIA	DLI	FSI
22	20	27

Table 24

Summary of Responses to tester Questionnaire

"Do you feel that the interviewing procedures (elicitation techniques, use of props, role-plays, etc.) you used during the study were the same as those you use in routine, day-to-day testing?"

SOMEWHAT	2
DEFINITELY	16

"Do you feel that the ratings you assigned during the study were, on the whole, more accurate, about as accurate, or less accurate than ratings you typically make in routine testing at your agency?"

NOT AS ACCURATE	1
ABOUT AS ACCURATE	17

"Do you feel that the accuracy of your ratings varied at certain times or points during the six-day testing period?"

NOT AT ALL	9
A LITTLE	6
SOMEWHAT	3

"Did you notice any differences in the composition of the examinee groups at the different agencies with respect to overall levels of proficiency, examinee reactions to interview techniques, etc.?"

NOT AT ALL	4
A LITTLE	4
SOMEWHAT	10

Table 24 (Continued)

"Do you feel that participation in the project was in any way interesting or beneficial to you?"

NOT AT ALL	1
SOMEWHAT	4
DEFINITELY	13